Welcome Message from the Scientific Program Committee Chair

AMIA 2016 Annual Symposium is the 40th anniversary of the Symposium. On behalf of the 2016 Scientific Program Committee we want this year to be a homecoming that energizes our new and longtime members, welcomes attendees who haven’t visited the meeting in a few years; and attracts first time attendees to the incredible informatics community AMIA represents.

AMIA is multi-disciplinary and interprofessional. We have the best of all worlds when it comes to soliciting and presenting the very best in informatics science and practice. In the rapidly moving world of health and healthcare, informatics innovation is integral to progress. No meeting captures the breadth and depth of informatics better than the AMIA Annual Symposium.

What’s New?

Diversified the SPC leadership and named four AMIA 2016 Vice Chairs: Noemie Elhadad, PhD; Mark Frisse, MD, MS, MBA, FACMI; Andrea Hartzler, PhD; and Madhu Reddy, PhD, FACMI
Unified the foundations and applications tracks to support work that crosses those boundaries
Appointed additional expert leaders to the SPC to manage the review process for more than 1200+ expected submissions
Updated the keywords to reflect the wide range of submitted content
Added option of three keywords per submission to improve the match between the submission topic and the expertise of the reviewers

The Symposium Experience

Sharing scientific discovery and demonstrating best practice and innovation is at the core of the AMIA Symposium experience. But the most unforgettable part of the meeting is the people. Informatics is a specialized field and the Symposium gathers our community of colleagues, peers, mentors, students and leaders for networking and fellowship, collaboration and friendship, and discussion and debate.

Wanda Pratt, PhD, FACMI
University of Washington
Chair, 2016 Scientific Program Committee
2016 AMIA Board of Directors

Officers

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

Blackford Middleton, MD, MPH, MSc, FACMI
Apertiva, Inc., and Harvard TH Chan School of Public Health
Past Chair

Titus K. Schleyer, DMD, PhD, FACMI
Treasurer

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

Directors

Wendy W. Chapman, PhD, FACMI
University of Utah

Theresa Cullen, MD, MS
Veterans Health Administration

Patricia Dykes, DNSc, MA, RN, FACMI
Brigham and Women’s Hospital

Peter J. Embi, MD, MS, FACMI
The Ohio State University

Cynthia S. Gadd, PhD, FACMI
Vanderbilt University

Gretchen Purcell Jackson, MD, PhD
Vanderbilt University Medical Center

Curtis P. Langlotz, MD, PhD, FACMI, FSIIM
Stanford University

Wanda Pratt, PhD, FACMI
University of Washington

Neil Sarkar, PhD, FACMI
Brown University
Dean F. Sittig, PhD, FACMI, FHIMSS
University of Texas Health Sciences Center at Houston
2013 - 2015

Jessica Tenenbaum, PhD
Duke Translational Medicine Institute

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

Ex-Officio Board Members

Suzanne Bakken, RN, PhD, FACMI
American College of Medical Informatics President
Columbia University
2015 - 2016

John T. Finnell, MD
Academic Forum Executive Committee Chair
Regenstrief Institute/Indiana University School of Medicine

Douglas B. Fridsma, MD, PhD, FACP, FACMI
President and CEO, AMIA

Michael Weiner, DO
Industry Advisory Council Chair
IBM

Laura K. Wiley, PhD
Student Working Group Representative
UC Denver Medical Center
AMIA 2016 Scientific Program Committee

Chair
Wanda Pratt, PhD
University of Washington

Vice Chairs
Noemie Elhadad, PhD
Columbia University

Mark Frisse, MD
Vanderbilt University

Andrea Hartzler, PhD
Group Health Research Institute

Madhu Reddy, PhD
Northwestern University

Members
Joanna Abraham, PhD
University of Illinois

Samantha Adams, PhD
Tilburg University

Jessica Ancker, PhD
Weill Cornell Medical College

Uba Backonja, PhD, RN
University of Washington

Sameer Badlani, MD
University of Chicago

Olivier Bodenreider, MD
National Library of Medicine

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

Dina Demner-Fushman, MD
National Library of Medicine

Jason Doctor, PhD
University of Southern California

David Dorr, MD
Oregon Health and Science University

Patricia Dykes, DNSc
Brigham and Women’s Hospital

Hamish S.F. Fraser, MBChB, MRCP, MSc
University of Leeds

Joydeep Ghosh, PhD
University of Texas Austin

David Hanauer, MD, MS
University of Michigan

Christopher Harle, PhD
Indiana University

Jina Huh, PhD
Michigan State University

Cathy Ivory, PhD
Vanderbilt University

Rebecca Crowley Jacobson, MD, MS
University of Pittsburgh

Bonnie Kaplan, PhD
Yale University

David Kaufman, PhD
Arizona State University

Abel Kho, MD
Northwestern University

Pedja Klasnja, PhD
University of Michigan

Paul Kleeberg, MD
Aldedade

Ross Koppel, PhD
University of Pennsylvania

Craig Kuziemsky, PhD
University of Ottawa

Brad Malin, PhD
Vanderbilt University

Benjamin Marlin
University of Massachusetts Amherst

Andrew Mellin, MD
McKesson

Genevieve Melton-Meaux, PhD
University of Minnesota

Casey Overby, PhD
University of Maryland School of Medicine

Adler Perotte, MD
Columbia University

Ari Pollack, MD
Seattle Children's Hospital/University of Washington

Guergana Savova, PhD
Children's Hospital Boston/Harvard Medical School

Katie Siek, PhD
Indiana University

David Sontag, PhD
New York University

Shane Stenner, MD
Vanderbilt University

Jimeng Sun, PhD
Georgia Institute of Technology

Tammy Toscos, PhD
Mirro Center for Research and Innovation

Tiffany Veinot, PhD
University of Michigan

Colin Walsh, MD, MA
Vanderbilt University

Chunhua Weng, PhD
Columbia University

Adam Wilcox, PhD
University of Washington

Lauren Wilcox, PhD
Georgia Institute of Technology

Meliha Yetisgen, PhD
University of Washington

Pierre Zweigenbaum, PhD
LIMSI-CNRS
<table>
<thead>
<tr>
<th>AMIA 2016 Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aakre, Christopher</td>
</tr>
<tr>
<td>Albers, David</td>
</tr>
<tr>
<td>Amyot PhD, Daniel</td>
</tr>
<tr>
<td>AAIAbdulsalam, Abdulrahman</td>
</tr>
<tr>
<td>Aldekhyyyel, Ranyah</td>
</tr>
<tr>
<td>Anazodo, Amechi N</td>
</tr>
<tr>
<td>Aarts, Jos</td>
</tr>
<tr>
<td>Alekseyenko, Alexander V</td>
</tr>
<tr>
<td>Ancker, Jessica S</td>
</tr>
<tr>
<td>Abar, Orhan</td>
</tr>
<tr>
<td>Alexander, Gregory L</td>
</tr>
<tr>
<td>Anders, Shilo H</td>
</tr>
<tr>
<td>Abbaszadegan, Hamed</td>
</tr>
<tr>
<td>Alexander, Gretchen B.</td>
</tr>
<tr>
<td>Andersen, Tariq O</td>
</tr>
<tr>
<td>Abdelrahman, Samir E</td>
</tr>
<tr>
<td>Alhammad, Ohoud Saad</td>
</tr>
<tr>
<td>Anganes, Amanda</td>
</tr>
<tr>
<td>Abdulai, Raolat</td>
</tr>
<tr>
<td>Alhuwail, Dari</td>
</tr>
<tr>
<td>Anjomshoaa, Hamideh</td>
</tr>
<tr>
<td>Abedtash, Hamed</td>
</tr>
<tr>
<td>Ali, Samira</td>
</tr>
<tr>
<td>Annechiarico, Robert P</td>
</tr>
<tr>
<td>Abrams, Denisha Nicole</td>
</tr>
<tr>
<td>Alipanah, Neda</td>
</tr>
<tr>
<td>Antani, Sameer</td>
</tr>
<tr>
<td>Acharya, Amit</td>
</tr>
<tr>
<td>Allen, Latricia</td>
</tr>
<tr>
<td>Anton, Bonnie</td>
</tr>
<tr>
<td>Adam, Terrence J</td>
</tr>
<tr>
<td>Almeida, Jonas S</td>
</tr>
<tr>
<td>Aphinyanaphongs, Yindalon</td>
</tr>
<tr>
<td>Adams, Meredith C. B.</td>
</tr>
<tr>
<td>Almeida, Mauricio B.</td>
</tr>
<tr>
<td>Ariosto, Deborah</td>
</tr>
<tr>
<td>Adamusiaik, Tomasz</td>
</tr>
<tr>
<td>Almoeen, Abdulgader</td>
</tr>
<tr>
<td>Arnold, Corey W</td>
</tr>
<tr>
<td>Afzal, Naveed</td>
</tr>
<tr>
<td>Alonso-Calvo, Raul</td>
</tr>
<tr>
<td>Arruda-Olson, Adelaide</td>
</tr>
<tr>
<td>Agu, Emmanuel</td>
</tr>
<tr>
<td>Alshehri, Majdah</td>
</tr>
<tr>
<td>Asare, Adam L</td>
</tr>
<tr>
<td>Ahmad, Suzan</td>
</tr>
<tr>
<td>Al-Taie, Zainab Salam</td>
</tr>
<tr>
<td>Asaro, Phillip</td>
</tr>
<tr>
<td>Ahmed, Waled Amed</td>
</tr>
<tr>
<td>Altamore, Rita</td>
</tr>
<tr>
<td>Ash, Nachman</td>
</tr>
<tr>
<td>Ahuja, Manik</td>
</tr>
<tr>
<td>Alves, Danielle Santos</td>
</tr>
<tr>
<td>Ashish, Naveen</td>
</tr>
<tr>
<td>Ahuja, Monika</td>
</tr>
<tr>
<td>Amante, Daniel</td>
</tr>
<tr>
<td>Atnoor, Deven</td>
</tr>
<tr>
<td>Airan-Javia, Subha L</td>
</tr>
<tr>
<td>Amato, Mary G</td>
</tr>
<tr>
<td>Aucoin, Jennifer M</td>
</tr>
<tr>
<td>Alaa, Ahmed Mohamed</td>
</tr>
<tr>
<td>Ambert, Kyle H.</td>
</tr>
<tr>
<td>Austin, Elizabeth</td>
</tr>
<tr>
<td>Alafaireet, Patricia E</td>
</tr>
<tr>
<td>Amith, Muhamamd Faheem</td>
</tr>
<tr>
<td>Austin, Robin</td>
</tr>
<tr>
<td>Alameddine, Sarah</td>
</tr>
<tr>
<td>Ammenwerth, Elske</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>Avali, Viji</td>
</tr>
<tr>
<td>Avrunin, George S.</td>
</tr>
<tr>
<td>Ayubi, Soleh</td>
</tr>
<tr>
<td>Azadmanjir, Zahra</td>
</tr>
<tr>
<td>Aziz, Ayesha</td>
</tr>
<tr>
<td>Azondekon, Roseric</td>
</tr>
<tr>
<td>Gbedegnon</td>
</tr>
<tr>
<td>Backonja, Uba</td>
</tr>
<tr>
<td>Badger, Martha Kimpton</td>
</tr>
<tr>
<td>Badji, Radja</td>
</tr>
<tr>
<td>Bai, Shasha</td>
</tr>
<tr>
<td>Bailey, Steffani R</td>
</tr>
<tr>
<td>Bakal, Gokhan</td>
</tr>
<tr>
<td>Baldwin, Jessica</td>
</tr>
<tr>
<td>Baldwin, Tyler</td>
</tr>
<tr>
<td>Banach, Mary</td>
</tr>
<tr>
<td>Bandyopadhyay PhD, Kakali</td>
</tr>
<tr>
<td>Banerjee, Ashis Gopal</td>
</tr>
<tr>
<td>Banger, Alison</td>
</tr>
<tr>
<td>Banning, Pamela</td>
</tr>
<tr>
<td>Bansal, Arvind</td>
</tr>
<tr>
<td>Bao, Riyue</td>
</tr>
<tr>
<td>Baptista, Roberto Silva</td>
</tr>
<tr>
<td>Barnes, Laura</td>
</tr>
<tr>
<td>Barnett, William</td>
</tr>
</tbody>
</table>
Bokov, Alex F
Boland, Mary Regina
Bona, Jonathan P
Bonomi, Luca
Borbolla, Damian Alberto
Boren, Suzanne A
Borland PhD, David
Botts, Nathan
Bouayad, Lina
Bouzillé, Guillaume
Bowles, Kathryn H
Boyce, Keith
Boyce, Richard David
Boyd, Andrew
Bradley, David James
Bragg, Mike
Braunstein, Mark L
Bray, Bruce E
Breene, Michael
Breitenstein, Matthew K.
Brennan, Caitlin W
Bressan, Nadja
Bright, Tiffani J
Britten, Tom
Brikey, Juliana J
Bronsburg Mr., Steve Edward
Brown, Stuart M
Brunson, Jason Cory
Buchan Jr., Kevin
Bucher, Brian Thomas
Buckel, Whitney
Burckhardt, Philipp
Burdick, Elisabeth
Burdick, Timothy E
Burke, Harry B.
Burke, Juandalyn Leaunda
Burlison, Jonathan
Burnett, Carl
Bustamante, Claudia
Byrd, Roy J
Byrne, Colene
Caballero Barajas, Karla Lizabeth
Caban, Jesus J
Cahan, Amos
Cai, Yi
Cairelli, Michael J.
Callahan, Alison
Callahan, Tiffany Jean Lee
Calvo, Richard Yee
Campbell, Robert C
Campbell, Terrance R
Campbell, Walter Scott
Campbell MD, James Richard
Campos, Manuel
Cantor, Michael
Caraballo, Pedro J
Carrington, Jane M.
Carroll, Mark
Carson, Kimberly J
Carson, Matthew
Carter, Ernest
Castro, Victor M
Cato RN, Kenrick
Ceusters, Werner
Chakrabarti, Shreya
Chakraborty, Swati
Champagne-Langabeer, Tiffany
Chandran, Uma
Chang, Hung-yang
Charney, Pam
Chartash, David
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Damal, Kavitha</td>
<td>Dermouche, Mohamed</td>
<td>Du, Fan</td>
</tr>
<tr>
<td>Daneshvari, Shamsi Rebecca</td>
<td>Dery, Samuel K K</td>
<td>Du, Shimeng</td>
</tr>
<tr>
<td>Daphtary, Kshama</td>
<td>Devarakonda, Murthy V</td>
<td>Du, Prerna</td>
</tr>
<tr>
<td>Darcy, Niamh</td>
<td>Dexheimer, Judith</td>
<td>Duan, Rui</td>
</tr>
<tr>
<td>Das MD, Anirudha</td>
<td>Dey, Sanjoy</td>
<td>Duggan, Dominic</td>
</tr>
<tr>
<td>Dasgupta, Dipanwita</td>
<td>Dhombres, Ferdinand</td>
<td>Duke, Jon</td>
</tr>
<tr>
<td>Davey, Adam</td>
<td>Dhopeshwarkar, Neil</td>
<td>Dullabh, Prashila</td>
</tr>
<tr>
<td>Davidson, Patricia</td>
<td>Diehl, Alexander D</td>
<td>Dumontier, Michel</td>
</tr>
<tr>
<td>Davies, Melanie</td>
<td>Dimitropoulos, Linda</td>
<td>Duncan, Jeffrey</td>
</tr>
<tr>
<td>Davis, Amanda</td>
<td>Dinakarpandian, Deendayal</td>
<td>Dunn, Adam G</td>
</tr>
<tr>
<td>Davis, Jesse</td>
<td>Ding, Ying</td>
<td>Dunn Lopez, Karen</td>
</tr>
<tr>
<td>Davis, Sharon E</td>
<td>Dinhofer, David S</td>
<td>Duran, Nelida</td>
</tr>
<tr>
<td>Day, Frank Cocchi</td>
<td>Divita, Guy</td>
<td>Dykes, Patricia C</td>
</tr>
<tr>
<td>Day, Michele E.</td>
<td>Dixon, Brian E</td>
<td>Dziadzko, Mikhail A</td>
</tr>
<tr>
<td>Dcruz, Jina J</td>
<td>Dligach, Dmitriy</td>
<td>Dzomba, Bari</td>
</tr>
<tr>
<td>de Bruijn, Berry</td>
<td>Doan, Raymond</td>
<td>Eckerle Mize, Dara</td>
</tr>
<tr>
<td>de Coronado, Sherri</td>
<td>Doan, Son</td>
<td>Edinger, Tracy</td>
</tr>
<tr>
<td>De Silva, Lalindra</td>
<td>Doberne, Julie W.</td>
<td>Edmunds, Margo</td>
</tr>
<tr>
<td>Dean, Nathan C</td>
<td>Doebbeling, Bradley N</td>
<td>Edwards, Todd</td>
</tr>
<tr>
<td>Del Fiol, Guilherme</td>
<td>Doing-Harris, Kristina Mary</td>
<td>Effken, Judith</td>
</tr>
<tr>
<td>Delaney, Connie White</td>
<td>Donevant, Sara</td>
<td>Eichmann, David</td>
</tr>
<tr>
<td>DeLaurentis PhD, Poching</td>
<td>Doshi, Riddhi</td>
<td>Eikey, Elizabeth Victoria</td>
</tr>
<tr>
<td>Deleger, Louise</td>
<td>Dowding, Dawn</td>
<td>Eisenberg, Floyd</td>
</tr>
<tr>
<td>Demiris, George</td>
<td>Driscoll, Patricia</td>
<td>Eisenstein, Eric</td>
</tr>
</tbody>
</table>
Galdzicki, Michal    Gichoya, Judy Wawira    Grabar, Natalia
Gallagher, Joel Louis    Gilbertson, David G    Grando, Maria Adela
Gallo De Moraes, Alice    Gilder, Jason    Grasso, Michael A.
Gallopyn, Guido    Giordano, Daniela    Gravenor, Stephanie
Galusca, Dragos    Girju, Roxana    Green, Amanda Constance
Gamache, Roland    Gligorijevic, Djordje    Griffis, Denis R
Garvin, Jennifer Hornung    Gobbel, Glenn T    Griffith, Janessa
Gaudioso, Carmelo    Goel, Anupam    Grose, Tim
Gaynor, Mark    Goel, Satyender    Grouin, Cyril
Gazarian, Priscilla K    Gold, Jeffrey A    Groza, Tudor
Geller, James    Gold, Rachel    Grundmeier, Robert Wayne
Gelzer MD, Andrea    Goldschmidt, Yaara    Guidry, Alicia
Geng, Yimin    Goldsmith, Denise    Gundlapalli, Adi V
Gennari, John H    Goldstein, Mary K.    Guo, Yufan
George, Ivan    Gong, Yang    Gupta, Anurag
Gershanik, Esteban    Gonna, Hanney    Gupta, Preyanshu
Gesner, Emily J    Goodwin, Rebecca Marci    Gurupur, Varadraj Prabhu
Gevaert, Olivier    Goodwin, Travis Reed    Gururaj, Anupama Edward
Ghassemi, Mohammad Mahdi    Gordon, Geoffrey D    Haas, Stephanie W
Ghebre, Yohannes T    Goswami, Chirayu    Hadji, Brahim
Ghiasvand, Omid    Gotz, David    Hagen, Mathew
Giannaris, Pericles Stavros    Gouripeddi, Ramkiran    Hagen, Michael D.
Giardina, Traber    Goyal, Dev    Haggstrom, David
Gibson, Bryan    Goyal, Nikhil    Hahn, Udo
<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hajagos, Janos</td>
<td>Hausam, Robert</td>
<td>Hillman, JoAnna L.</td>
</tr>
<tr>
<td>Hall, Amanda K</td>
<td>Hays, David</td>
<td>Hirsch, Tad</td>
</tr>
<tr>
<td>Hall, Eric S</td>
<td>Hazen, Rebecca J</td>
<td>Hirschman, Lynette</td>
</tr>
<tr>
<td>Hall, Leslie Kelly</td>
<td>He, Yongqun</td>
<td>Hixson, John David</td>
</tr>
<tr>
<td>Hameed, Farhan</td>
<td>He, Yuqi</td>
<td>Ho, King Chung</td>
</tr>
<tr>
<td>Hammer MD, Richard</td>
<td>He, Ze</td>
<td>Hochheiser, Harry</td>
</tr>
<tr>
<td>Hamon, Thierry</td>
<td>He, Zhe</td>
<td>Hoffman, James M</td>
</tr>
<tr>
<td>Han, Sifei</td>
<td>He PhD, Shan</td>
<td>Hoffman, Mark</td>
</tr>
<tr>
<td>Hanauer, David</td>
<td>Heard, Kevin M</td>
<td>Hogan, Linda</td>
</tr>
<tr>
<td>Harle, Christopher</td>
<td>Hebert, Courtney Lee</td>
<td>Hogan, William</td>
</tr>
<tr>
<td>Harper, Jeremy</td>
<td>Heermann Langford, Laura K</td>
<td>Holl, Felix</td>
</tr>
<tr>
<td>Harper, Marvin</td>
<td>Heinrich, Julia N</td>
<td>Hollberg, Julie</td>
</tr>
<tr>
<td>Harrell, Morgan</td>
<td>Hekmatnejad, Mohammad</td>
<td>Holmes, John H.</td>
</tr>
<tr>
<td>Harrington, Linda</td>
<td>Helmer, Karl</td>
<td>Homer, Mark L</td>
</tr>
<tr>
<td>Harris, Kimberly</td>
<td>Hempelmann, Christian F.</td>
<td>Hongsermeier, Tonya</td>
</tr>
<tr>
<td>Harris, Kimberly A</td>
<td>Henao, Ricardo</td>
<td>Hoonakker, Peter</td>
</tr>
<tr>
<td>Harris, Marcelline</td>
<td>Henderson, Jette</td>
<td>Hoot, Nathan R</td>
</tr>
<tr>
<td>Harrison, Andrew M.</td>
<td>Henriksson, Aron</td>
<td>Horne, Benjamin D</td>
</tr>
<tr>
<td>Hart, Chip</td>
<td>Her, Qoua Liang</td>
<td>Horridge, Matthew</td>
</tr>
<tr>
<td>Hartman, Terry Scott</td>
<td>Hernandez Caffot, Susana</td>
<td>Hors-Fraile, Santiago</td>
</tr>
<tr>
<td>Hartzler, Andrea L.</td>
<td>Herr, Timothy M</td>
<td>Horsky, Jan</td>
</tr>
<tr>
<td>Hasan, Sadid A.</td>
<td>Hersh, William R</td>
<td>Hosseini, Narges</td>
</tr>
<tr>
<td>Hassanzadeh, Oktie</td>
<td>Herzig, Carolyn T.A.</td>
<td>Hovenga, Evelyn JS</td>
</tr>
<tr>
<td>Haug, Peter John</td>
<td>Hickman, Timothy P</td>
<td>Hoxha, Julia</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Hoyt, Robert E</td>
<td>Iyengar, Sriram</td>
<td>Johnson, Sharon Ann</td>
</tr>
<tr>
<td>Hribar, Michelle R</td>
<td>Jackson, Gretchen Purcell</td>
<td>Johnson, Stephen Bennett</td>
</tr>
<tr>
<td>Hruby, Gregory William</td>
<td>Jacobs, Jason</td>
<td>Johnson, Steven G</td>
</tr>
<tr>
<td>Hsu, Chun-Nan</td>
<td>Jacobs, Maia</td>
<td>Jones, Josette F.</td>
</tr>
<tr>
<td>Hsu, William</td>
<td>Jacobs, Robin J.</td>
<td>Joseph, Joshua W</td>
</tr>
<tr>
<td>Hu, Jianying</td>
<td>Jacobs, Tom</td>
<td>Kaelber, David</td>
</tr>
<tr>
<td>Hu, Lu</td>
<td>Jadhav, Ashutosh</td>
<td>Kahn Jr, Charles E.</td>
</tr>
<tr>
<td>Hu, Peter</td>
<td>Jain, Sanjay R</td>
<td>Kale, David</td>
</tr>
<tr>
<td>Hu, Zhen</td>
<td>Jain, Viral</td>
<td>Kalenderian, Elsbeth</td>
</tr>
<tr>
<td>Huang, Chih-Wei</td>
<td>James, Andrew</td>
<td>Kang, Hong</td>
</tr>
<tr>
<td>Hudak, Christine Angela</td>
<td>Jamieson, Trevor</td>
<td>Kang, Tian</td>
</tr>
<tr>
<td>Hughes, Kevin S</td>
<td>Jamoom, Eric</td>
<td>Kanga, Samuel Gichohi</td>
</tr>
<tr>
<td>Hulse, Nathan C</td>
<td>Jelen, Benjamin</td>
<td>Kannampallil, Thomas G</td>
</tr>
<tr>
<td>Hultman, Gretchen M</td>
<td>Jenders, Robert Allen</td>
<td>Kanter, Andrew Stuart</td>
</tr>
<tr>
<td>Hurdle, John F</td>
<td>Ji, Xiaonan</td>
<td>Kapoor, Akshat</td>
</tr>
<tr>
<td>Hurley, Ann</td>
<td>Jiang, Guoqian</td>
<td>Karami, Amir</td>
</tr>
<tr>
<td>Hussaini, Khaleel</td>
<td>Jiang, Ling</td>
<td>Karavite, Dean J</td>
</tr>
<tr>
<td>Hylock, Ray Hales</td>
<td>Jiang, Xiaoqian</td>
<td>Karipineni, Neelima</td>
</tr>
<tr>
<td>Hyun, Sookyung</td>
<td>Jimeno Yepes, Antonio</td>
<td>Karlsson, Daniel</td>
</tr>
<tr>
<td>Imler, Timothy David</td>
<td>Jing, Xia</td>
<td>Kartoun, Uri</td>
</tr>
<tr>
<td>Iribarren, Sarah</td>
<td>Joe, Jonathan</td>
<td>Kassakian, Steven Z</td>
</tr>
<tr>
<td>Islam, Roosan</td>
<td>John, Rita Marie</td>
<td>Kasthurirathne, Suranga Nath</td>
</tr>
<tr>
<td>Ito, Márcia</td>
<td>Johnson, Constance M</td>
<td>Kate, Rohit</td>
</tr>
<tr>
<td>Ivory, Catherine H</td>
<td>Johnson, Patricia</td>
<td>Katsura, Chris</td>
</tr>
<tr>
<td>Kaufman, David R</td>
<td>Kim, Tae Youn</td>
<td>Koss, Michele</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Kavuluru, Ramakanth</td>
<td>Kimura, Eizen</td>
<td>Kothapeta, Vishnoo Charan Reddy</td>
</tr>
<tr>
<td>Kayaalp, Mehmet</td>
<td>King, Andrew J</td>
<td>Kotov, Alexander</td>
</tr>
<tr>
<td>Kelley, Marjorie M</td>
<td>Kirbiyik, Uzay</td>
<td>Kouril, Michal</td>
</tr>
<tr>
<td>Kelly, Joseph</td>
<td>Kirchner, Lester</td>
<td>Krall, Michael</td>
</tr>
<tr>
<td>Kennell Jr., Timothy Irving</td>
<td>Kite, Bobbie</td>
<td>Krause, Josua</td>
</tr>
<tr>
<td>Kerley, Denise</td>
<td>Kitsiou, Spyros</td>
<td>Krause 3463851, Tim</td>
</tr>
<tr>
<td>Ketchum, Andrea</td>
<td>Klann, Jeffrey Gordon</td>
<td>Krauss, Deborah</td>
</tr>
<tr>
<td>Khalifa, Aly MA</td>
<td>Klassen, Prescott</td>
<td>Krishnaswamy, Asha</td>
</tr>
<tr>
<td>Khalilia, Mohammed</td>
<td>Klein, Krystal A</td>
<td>Krive, Jacob</td>
</tr>
<tr>
<td>Khan, Faiza</td>
<td>Kleinberg, Samantha</td>
<td>Kroth, Philip J</td>
</tr>
<tr>
<td>Khan, Naqi</td>
<td>Klimov, Denis</td>
<td>Kuhn, Thomson M</td>
</tr>
<tr>
<td>Khare, Ritu</td>
<td>Kneale, Laura</td>
<td>Kulanthaivel, Anand</td>
</tr>
<tr>
<td>Kharrazi, Hadi</td>
<td>Knecht, Karen</td>
<td>Kulikowski, Casimir A.</td>
</tr>
<tr>
<td>Kho, Abel N</td>
<td>Knoll, Benjamin C</td>
<td>Kumar, Aman</td>
</tr>
<tr>
<td>Khunlertkit, Adjhaporn Nana</td>
<td>Ko, Hon Sum</td>
<td>Kumar, Manish</td>
</tr>
<tr>
<td>Khurana, Himica</td>
<td>Kohlmayer, Florian</td>
<td>Kunisch, Joseph M</td>
</tr>
<tr>
<td>Kilicoglu, Halil</td>
<td>Kohn, Martin S</td>
<td>Kuo, Tsung-Ting</td>
</tr>
<tr>
<td>Killeen, James P</td>
<td>Kolh, Philippe</td>
<td>Kury, Fabricio Sampaio Peres</td>
</tr>
<tr>
<td>Kim, Christopher Y</td>
<td>Komandur Elayavilli, Ravikumar</td>
<td>Kuttler, Kathryn Gibb</td>
</tr>
<tr>
<td>Kim, Katherine K</td>
<td>Koola, Jejo David</td>
<td>Kuziemsky, Craig</td>
</tr>
<tr>
<td>Kim, Min-hyung</td>
<td>Korach, Zfania Tom</td>
<td>Labkoff, Steven E</td>
</tr>
<tr>
<td>Kim, Sujin</td>
<td>Koralkar, Rajesh H</td>
<td>Lacson, Ronilda</td>
</tr>
<tr>
<td>Kim, Sun</td>
<td>Koru, Gunes</td>
<td>Lai, Kenneth H</td>
</tr>
<tr>
<td>Name</td>
<td>First Name</td>
<td>Last Name</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Lake, Erica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lalley, Cathy R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamas, Eugenia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landman, Adam B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang, Lisa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Langlotz, Curtis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leader, Joseph B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ledbetter, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee, Eva K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee, Jaehoon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee, John</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee, Mikyoung</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee, Yungyung</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lehmann, Harold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lei, Victor John</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lele, Omkar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leroy, Gondy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesh, Kathy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lessard PhD, Lysanne</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesselroth, Blake Justin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu, Mike</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leung, Tiffany I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leviss, Jonathan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levy, Bruce</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lorenzi, Virginia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Dingcheng</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Haiquan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Man</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Peter W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Ying</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liang, Chen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liaw, Siaw-Teng</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lichtner, Valentina</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likumahuwa-Ackman, Sonja</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lim Choi Keung, Sarah N.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin, Yu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lindquist, Richard R.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lingren, Todd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listhaus, Alyson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Chuanren</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Haifeng</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Hongfang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Jie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Leslie S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Ruiling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu, Yang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lockett-Brown, Linda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loonsk, John W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lorberbaum, Tal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lou, Jennie Q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lourie, Eli M.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lovely, Jenna K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowery-North, Douglas W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lu, Zhiyong</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lucero, Robert James</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lucic, Ana</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luna, Jorge Marcial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luo, Yuan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luyckx, Kim</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyalin, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyalil, David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyle, Jay</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyles, Courtney</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyman, Jason A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyon, Lawrence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ma, Sisi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mabotuwana, Thusitha</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MacLeod, Haley</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madani, Sina</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madkour, Mohcine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magee, Daniel J</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magge, Arjun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magrabi, Farah</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mahajan, Satish M</td>
<td>Marshall, Robert Carter</td>
<td>McInnes, D. Keith</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Mahnke, Andrea</td>
<td>Marsolo, Keith</td>
<td>McKanna, James</td>
</tr>
<tr>
<td>Mahoui, Malika</td>
<td>Martial, Marc-Aurel</td>
<td>McNamara, Mary</td>
</tr>
<tr>
<td>Major, Vincent</td>
<td>Martinez, Gabriela</td>
<td>McPeek Hinz, Eugenia R</td>
</tr>
<tr>
<td>Malec, Scott Alexander</td>
<td>Martinez-Romero, Marcos</td>
<td>Medlock, Stephanie</td>
</tr>
<tr>
<td>Malin, Bradley</td>
<td>Martin-Sanchez, Fernando J</td>
<td>Meehan, Rebecca</td>
</tr>
<tr>
<td>Malmasi, Shervin</td>
<td>Massoudi, Barbara L</td>
<td>Meeker, Daniella</td>
</tr>
<tr>
<td>Mamykina, Lena</td>
<td>Matheny, Michael Edwin</td>
<td>Mehta, Prachi</td>
</tr>
<tr>
<td>Manaktala, Sharad</td>
<td>Matney, Susan A.</td>
<td>Melton, Brittany</td>
</tr>
<tr>
<td>Manard, William Troy</td>
<td>Maviglia, Saverio M</td>
<td>Menon, Prahlad G</td>
</tr>
<tr>
<td>Mancuso, Mary Patricia</td>
<td>Maygers, Joyce</td>
<td>Mentis, Helena M</td>
</tr>
<tr>
<td>Mandel, Joshua C</td>
<td>Mayne, Stephanie L</td>
<td>Merrill, Jacqueline</td>
</tr>
<tr>
<td>Mandl, Kenneth D</td>
<td>Mays, Eric</td>
<td>Metzger, Marie Helene</td>
</tr>
<tr>
<td>Mani, Subramani</td>
<td>McCallie Jr, David</td>
<td>Meystre, Stephane</td>
</tr>
<tr>
<td>Manion, Frank J</td>
<td>McClay, James C</td>
<td>Mezitis, Nicholas H.E.</td>
</tr>
<tr>
<td>Mankowitz, Scott</td>
<td>McCoy Jr, Thomas H</td>
<td>Michalkiewicz, Michael</td>
</tr>
<tr>
<td>Manning, John D</td>
<td>McDonald, Clement J</td>
<td>Michalowski, Martin</td>
</tr>
<tr>
<td>Manos, Eva LaVerne</td>
<td>McFarlane, Timothy D</td>
<td>Michalowski, Wojtek</td>
</tr>
<tr>
<td>Maram, Naveen</td>
<td>McGlothlin, James P</td>
<td>Michel, Jeremy</td>
</tr>
<tr>
<td>Marc, David T</td>
<td>McGrath, Scott Patrick</td>
<td>Middleton, Blackford</td>
</tr>
<tr>
<td>Marcu, Gabriela</td>
<td>McGreevey III, John D</td>
<td>Mikles, Sean</td>
</tr>
<tr>
<td>Markey, Mia K</td>
<td>McGregor, Kyle A</td>
<td>Miksch, Tim</td>
</tr>
<tr>
<td>Marlin, Benjamin M</td>
<td>McGregor, Kyle Aaron</td>
<td>Miles, Stephen</td>
</tr>
<tr>
<td>Marquard, Jenna</td>
<td>McInnes, Bridget</td>
<td>Militello, Laura G</td>
</tr>
</tbody>
</table>
Miller, Andrew D  Morey, Jose  Nakikj, Drashko
Miller, Christopher M  Morgan, Stephen James  Nandigam, Hari Krishna
Miller, Timothy A  Mork, James G  Narus, Scott P
Millery, Mari  Mork, Peter  Natarajan, Annamalai
Min, Hua  Morlane-Hondere, François  Naumann, Tristan
Minich, Thomas  Mosa, Abu Saleh Mohammad  Nazi, Kim Marie
Minnier, Jessica  Moss, Laura  Nebeker, Jonathan Rich
Miotto, Riccardo  Motiwala, Tasneem  Nelson, Stuart James
Mishra, Rashmi  Mougin, Fleur  Neveol, Aurelie
Mitchell, Jayne  Mower, Justin  Newbold, Susan K
Mitchell, Terri H  Mowery, Danielle  Ng, Kenney
Mittu, Ranjeev  Mrabet, Yassine  Ngufor, Che
Mlaver, Eli  Mudd, James O.  Nguyen, Anthony
Mo, Huan  Mundt, Barbie J  Nguyen, Duc
Modave, François  Munson, Sean  Nguyen, Lien B.
Moen, Anne  Murcko, Anita Cecilia  Nguyen, Vickie
Mohammad, Amir  Murphy MD, Robert Eugene  Ni, Yizhao
Mohan, Santosh  Musser, R. Clayton  Nielson, Jeffrey Arthur
Mohan, Vishnu  Myerburg, Stuart  Ning, Xia
Moidu, Khalid  Myers, Risa  Nordo, Amy Harris
Mollah, Shamim  Myneni, Sahiti  Nosal, Sarah C
Monte, Robert  Nagarajan, Radhakrishnan  Novak, Laurie Lovett
Moore, Carlton Reid  Nahm, Eun-Shim  Nunes, Fátima L S
Morales, Miraida  Nakayama, Masaharu  Obara, Piotr
Obradovic, Zoran
O'Bryan, Kevin
Ochs, Christopher
Odlum, Michelle
Ogallo, William
Ogunyemi, Omolola
OHoro, John C
O'Leary, Kathleen
Oliveira, Eliel
Olsen, Joanne
Onega, Tracy
Opoku-Boateng, Gloria A
Orreggio, Giordi
Osterman, Travis
Otolorin, Abiodun
Overgaard, Shauna
Overhage, J. Marc
Ozaydin, Bunyamin
Ozery-Flato, Michal
Ozkaynak, Mustafa
Pacheco, Jennifer Allen
Painter, Ian
Palchuk, Matvey B
Palen, Ted E
Pan, Eric Chungren
Pan, Xuequn
Papana Dagiasis, Ariadni
Park, Albert
Park, Anne
Park, Hyeoun-Ae
Park, Jung In
Parsia, Bijan
Pasupathy, Kalyan
Patel, Aditi G.M.
Patel, Vimla L
Patel, Vipul
Patil, Vaibhav
Patrick PhD, Timothy B
Patrick, Jon D
Patterson, Olga V
Payne, Philip
Pedersen, Jay
Peek, Niels
Peeples, Malinda
Pell, Jonathan
Pepper, Catherine
Perl, Yehoshua
Perotte, Adler
Petrova, Mila
Petzel, Sue v
Pfeffer, Michael Adam
Pfiffner, Pascal Bruno
Phillips, Andrew
Phillips, Mark H
Piscotty Jr., Ronald John
Pivovarov, Rimma
Plasek, Joseph M
Platt, Jodyn
Poikonen, John
Pollack, Ari
Polzin, Thomas
Poole, Sarah
Poon, Hoifung
Popescu, Mihail
Popovic, Jennifer R
Porat, Talya
Posada, Jose David
Posteraro, Robert H
Poterack, Karl
Pozzi, Giuseppe
Prager, John M.
Pranaat, Robert
Prasser, Fabian
Rule, Adam
Rumshisky, Anna
Runger, George
Rusanov, Alex
Rusin, Craig G
Rusincovitch, Shelley A
Russ, Alissa L
Ruttenberg, Alan
Sacchi, Lucia
Sadasivam, Rajani S
Sadoghi, Mohammad
Safarova, Mayya S
Safran, Charles
Sahoo, Satya S
Salmasian, Hojjat
Saltz, Joel
Saltz, Mary
Samet, Saeed
Sandor, Nicolae Lucian
Sankaran, Sharlini
Sarantis, Demetrios
Sarcevic, Aleksandra
Sarkar, Indra Neil
Sarkar, Urmimala
Sarker, Abeed
Sarma, Karthik
Venkataraman
Satz, Wayne
Saupp, Christi
Savova, Guergana
Sawesi, Suhila A
Schaefbauer, Christopher
Schaffer, Jonathan Lawrence
Schaub, Marc Andreas
Scherpbier, Harm
Schlegel, Daniel R
Schleiden, Loren James
Schmaltz, Chester Lee
Schnall, Rebecca
Schneider, David F
Schneider, Jodi
Schreiber, Richard
Schubert, Carolyn
Schwei, Kelsey M
Scotch, Matthew
Seale, Deborah Elaine
Segal, Mark
Selva, Thomas
Senathirajah, Yalini
Seppälä, Selja
Seymour, Rachel
Shachak, Aviv
Shahar, Yuval
Shang, Howard
Shang, Ning
Shapiro, Jason S
Shapiro, Michael R.
Sharma, Ashish
Sharma, Deepak K
Shatkay, Hagit
Shea, Christopher M
Shea, Kimberly
Sheets, Lincoln
Shelov, Eric D
Shen, Shiwen
Shenson, Jared Andrew
Shenvi, Edna C
Shi, Jianlin
Shi, Lingyun
Shi, Wei
Shimada, Stephanie L.
Shimoyama, Mary
Shivade, Chaitanya
<table>
<thead>
<tr>
<th>Last Name, First Name</th>
<th>First Name, Last Name</th>
<th>First Name, Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shojaee Bakhtiari, Ali</td>
<td>Snowdon, Jane</td>
<td>St. Jacques, Paul</td>
</tr>
<tr>
<td>Shyr, Yu</td>
<td>Sobko, Heather J</td>
<td>Staes, Catherine J</td>
</tr>
<tr>
<td>Siek, Katie A</td>
<td>Sockolow, Paulina</td>
<td>Starren, Justin</td>
</tr>
<tr>
<td>Silveira, Patricia C</td>
<td>Sohn, Sunghwan</td>
<td>Steigman, Michael R</td>
</tr>
<tr>
<td>Simms, Andrew</td>
<td>Solomonides, Tony</td>
<td>Stepanek, Richard</td>
</tr>
<tr>
<td>Simon, Gyorgy</td>
<td>Soman, Sandeep</td>
<td>Stephan, Christina L</td>
</tr>
<tr>
<td>Sims, Maria</td>
<td>Song, Yiqiang</td>
<td>Stoddard, Gregory J</td>
</tr>
<tr>
<td>Sims, Shannon</td>
<td>Sontag, David</td>
<td>Stone III, Norman</td>
</tr>
<tr>
<td>Sincan, Murat</td>
<td>Sosic, Rok</td>
<td>Strasberg, Howard R</td>
</tr>
<tr>
<td>Singh, Karandeep</td>
<td>Soto-Campos, Gerardo</td>
<td>Su, Hang</td>
</tr>
<tr>
<td>Singh, Siddharth</td>
<td>Soualmia, Lina F</td>
<td>Suh, Hyewon</td>
</tr>
<tr>
<td>Sissung, Tristan</td>
<td>South, Brett Ray</td>
<td>Sujansky, Walter</td>
</tr>
<tr>
<td>Sittig, Dean F</td>
<td>Soysal, Ergin</td>
<td>Sulieman, Lina</td>
</tr>
<tr>
<td>Skapik, Julia</td>
<td>Spallek, Heiko</td>
<td>Sumner II, Walton</td>
</tr>
<tr>
<td>Skiba, Diane J</td>
<td>Speier, William</td>
<td>Sun, Simon</td>
</tr>
<tr>
<td>Skipper, Jamie</td>
<td>Speroff, Theodore</td>
<td>Sun, Zhaonan</td>
</tr>
<tr>
<td>Sklenar, Jiri</td>
<td>Sperrin, Matthew</td>
<td>Sutton, Joseph A</td>
</tr>
<tr>
<td>Slager MS, Stacey</td>
<td>Spisla, Christine</td>
<td>Sward, Katherine A</td>
</tr>
<tr>
<td>Slight, Sarah Patricia</td>
<td>Spooner, Stephen Andrew</td>
<td>Swartz, Jordan</td>
</tr>
<tr>
<td>Smail, Craig</td>
<td>Spyropoulos, Basile P.</td>
<td>Sweet, Micheal</td>
</tr>
<tr>
<td>Smalheiser, Neil R.</td>
<td>Sridharan, Vishnupriya</td>
<td>Sylvestre, Emmanuelle</td>
</tr>
<tr>
<td>Smith, Ann DeBord</td>
<td>Srinivas, Preethi</td>
<td>Taft, Teresa</td>
</tr>
<tr>
<td>Smith, Catherine Arnott</td>
<td>Srinivasan, Arunkumar</td>
<td>Takesue, Blaine</td>
</tr>
<tr>
<td>Smith, Wade P</td>
<td>Srinivasan, Padmini</td>
<td>Talbert, Douglas</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Wang, Chunye</td>
<td>Weiner, Mark</td>
<td>Wojtusiak, Janusz</td>
</tr>
<tr>
<td>Wang, Dongwen</td>
<td>Weir, Charlene</td>
<td>Wong, David C</td>
</tr>
<tr>
<td>Wang, Fei</td>
<td>weisenthal, samuel</td>
<td>Wood, Grant M</td>
</tr>
<tr>
<td>Wang, Haijun</td>
<td>Weiskopf, Nicole Gray</td>
<td>Woodcock, Deborah Virginia</td>
</tr>
<tr>
<td>Wang, Lucy Lu</td>
<td>Weiss, Jeremy</td>
<td>Woollen, Janet</td>
</tr>
<tr>
<td>Wang, Meng</td>
<td>Welch, Brandon M</td>
<td>Workman, Terri Elizabeth</td>
</tr>
<tr>
<td>Wang, Shuang</td>
<td>Wells, Kristen J.</td>
<td>Wright, Adam</td>
</tr>
<tr>
<td>Wang, Shunchao</td>
<td>Welty, Christopher A</td>
<td>Wright, Julie A</td>
</tr>
<tr>
<td>Wang, Xiang</td>
<td>Weng, Chunhua</td>
<td>Wu, Cai</td>
</tr>
<tr>
<td>Wang, Xiaoyan</td>
<td>West, Vivian L</td>
<td>Wu, Danny TY</td>
</tr>
<tr>
<td>Wang, Yajuan</td>
<td>Westra, Bonnie L</td>
<td>Wu, Han</td>
</tr>
<tr>
<td>Wang, Yan</td>
<td>White, Andrew</td>
<td>Wu, Stephen</td>
</tr>
<tr>
<td>Wang, Yanshan</td>
<td>White, Peggy</td>
<td>Wu, Yonghui</td>
</tr>
<tr>
<td>Wang, Yue</td>
<td>Whiting, Kandace Mikel</td>
<td>Wu, Zhaoming</td>
</tr>
<tr>
<td>Warner, Jeremy L.</td>
<td>Wickramasinghe, Nilmini</td>
<td>Xia, Fei</td>
</tr>
<tr>
<td>Warner, Phillip B</td>
<td>Wilbanks, Bryan Anthony</td>
<td>Xia, Zongqi</td>
</tr>
<tr>
<td>Weatherall, James</td>
<td>Wilkinson, Katie</td>
<td>Xie, Guotong</td>
</tr>
<tr>
<td>Weaver, Charlotte Ann</td>
<td>Willett, DuWayne Lee</td>
<td>Xie, Wei</td>
</tr>
<tr>
<td>Weber, Jens</td>
<td>Williams, Marc S</td>
<td>Xie, Yue</td>
</tr>
<tr>
<td>Wei, Chih-Hsuan</td>
<td>Williams, Richard</td>
<td>Xiong, Wenting</td>
</tr>
<tr>
<td>Wei, Duo (Helen)</td>
<td>Wilson, Marisa L</td>
<td>Xu, Yaomin</td>
</tr>
<tr>
<td>Wei, Wei</td>
<td>Winden, Tamara</td>
<td>Yakimischak, David</td>
</tr>
<tr>
<td>Wei, Xin</td>
<td>Windle, John R</td>
<td>Yale, Ken</td>
</tr>
<tr>
<td>Weibel, Nadir</td>
<td>Woeltje, Keith F</td>
<td>Yan, Chao</td>
</tr>
</tbody>
</table>
Yang, Shiming

Yang, Yushi

Yanover, Chen

Yarbrough, William Claiborne

Ye, Cheng

Yen, Po-Yin

Yi, Haeseung

Yim, Wen-wai

Yorks, Melissa

Yoshida, Eileen

Yu, Denny

Yu, Sheng

Yu Jr., Feliciano Buenviaje

Yusuf, Rafeek A

Zachary, Iris

Zaleski, John

Zangwill, Linda

Zech, John R.

Zeng, Jia

Zhang, Danchen

Zhang, Guo-Qiang

Zhang, Hao Howard

Zhang, Jinghe

Zhang, Linda

Zhang, Ping

Zhang, Rui

Zhang, Shaodian

Zhang, Songmao

Zhang, Yi

Zhang, Yiye

Zhang, Yongkang

Zhang, Zhe

Zhao, Yiqing

Zhen, Yi

Zheng, Hua

Zheng, Jiaping

Zheng, Shuai

Zhu, Haining

Zielinski, Gail Marie

Zimmermann, Michael T

Zolnoori, Maryam

Zorc, Joseph

Zuckerman, Alan E

Zweigenbaum, Pierre
AMIA 2016 Distinguished Reviewers

Madjar Alshehri
Barbara Berkowitz
Terrance Campbell
Eun-Kyoung Choe
George Demiris
Karen Dunn Lopez
Scott Evans
Anthony Faiola
Amy Franklin
Roxana Girju
Rebecca Goodwin
Jan Horsky
Maia Jacobs

Ben Jelen
Thomas Kannampallil
Gondy Leroy
Sonja Likumahuwa-Ackman
Laura Militello
Jonathan Nebeker
Dusan Ramljak
Alex Rosenthal
Aleksandra Sarcevic
Christopher Schaefbauer
Yuval Shohar
Charlene Weir
NOTICE

Medicine is an ever-changing science. As new research and clinical experience broaden our knowledge, changes in treatment and drug therapy are required. The authors and the publishers of this work have checked with sources believed to be reliable in their efforts to provide information that is complete and generally in accord with the standards accepted at the time of publication. However, in view of the possibility of human error or changes in medical sciences, neither the authors nor the publisher nor any other party who has been involved in the preparation or publication of this work warrants that the information contained herein is in every respect accurate or complete, and they are not responsible for any errors or omissions or for the results obtained from use of such information. Readers are encouraged to confirm the information contained herein with other sources. For example and in particular, readers are advised to check the product information sheet included in the package of each drug they plan to administer to be certain that the information contained in this book is accurate and that changes have not been made in recommended dose or in the contraindication for administration. This recommendation is of particular importance in connection with new or infrequently used drugs.
Obtaining a timeline of clinical events described in clinical narratives, even if only classified as historical or current, is very useful for caregivers. To automatically classify clinical events found in clinical notes as happening before, after, or overlapping with the clinical note creation time, we compared a selection of classification machine learning algorithms with a baseline. The highest accuracy measured reached 84.2%.

Introduction

Clinical narratives are rich source of information about the patient’s clinical history, progress and treatment plans marking major clinical events that occur over the course of longitudinal care. Oftentimes, it is crucial for caregivers to obtain a summarized overview of these clinical events, for example in the form of a visual timeline, to assist with making accurate care decisions. By classifying each clinical event mentioned in a clinical document with respect to the document creation time, we can obtain a high-level summarized timeline. These clinical events can be classified into one of four temporal relations to the document creation time: BEFORE (historical events that strictly ended before the document creation time), AFTER (future events that will start after the document creation time), OVERLAP (current events that hold true during the document creation time) and BEFORE/OVERLAP (events that started before the document creation time and continue to the present). Using automated natural language processing techniques combined with supervised machine learning, we evaluate the performance of a wide selection of algorithms and discuss the impact of features and parameters fine-tuning on the final classification accuracy.

Method

The THYME (Temporal Histories of Your Medical Events) corpus for colon cancer patients from Mayo clinic was used in this study. In the context of the 2015 Clinical TempEval competition, each clinical note in the corpus was manually annotated with clinical events and their temporal relations to the document creation time. The corpus was randomly divided into a training set (n=293) and a testing set (n=147). To automatically classify each clinical event temporal relation with the document creation time, we first processed each clinical note with cTAKES to extract lexical, morphological and syntactic features, as shown in Table 1. These features were augmented with additional event attributes (Polarity, Modality, Degree and Type) also provided in the reference standard. Features listed in Table 1 were then extracted for each word annotated as a clinical event in the reference standard, as well as for their surrounding context (local context window of 5 preceding and 5 following words, lemmas, parts of speech tags, and chunks). The Python scikit-learn machine learning toolkit was used with selected number of machine learning algorithms: linear support vector classifier (LinearSVC in scikit-learn), decision tree (DecisionTreeClassifier), random forest (RandomForestClassifier), stochastic gradient descent (SGDClassifier), and naïve Bayes (MultinomialNB). The CRF++ tool was utilized to build a sequence labeling model based on conditional random fields algorithm. All algorithms were trained on the training set and evaluated on the testing set. As baseline, we used the most frequent temporal relation category (i.e., OVERLAP) for all relations. The evaluation metric was accuracy (number of correctly classified instances divided by the total number of instances). Note that accuracy is equivalent to the F1-measure in this case, since all instances must be classified into one of four possible labels.

Results

When evaluated with the held-out testing set, accuracy ranged from 74.4% to 84.2% (Table 2). In general, algorithms performed significantly better when including features from their local context (e.g., SGDClassifier), but a few had a lower or no improvement in accuracy with these features (e.g., RandomForestClassifier). The support
vector machine (LinearSVC) cost of classification parameter $C$ is used to control the complexity of the model and its tendency to overfit (a higher value indicates more overfitting but less error on the training set). Using a grid search algorithm (GridSearchCV) to find the best $C$ value can have determinant effects on algorithm performance, as evidenced when setting $C=0.024$ for the LinearSVC algorithm; which increased accuracy from 81.8% to 84.2%. Similarly, performing grid search for the $\alpha$ parameter in a naïve Bayes (MultinomialNB) algorithm allowed for improved accuracy. The decision tree (DecisionTreeClassifier) algorithm did not benefit from local context features, and the random forest (RandomForestClassifier) performance was even penalized. The latter could be due to the random selection strategy of features that are used to train decision trees on sub-samples from the training data. The final classification is performed by averaging the output labels from the decisions trees in the forest. When poor features are randomly selected for some decision trees in the forest, they tend to affect the final classification decision. When using a feature selection limited to the top 10000 features, the random forest classification improves (indicated with * in Table 2), indicating that this algorithm is sensitive to the feature sets used for training. Note that a sequence labeling classification approach with CRF++ attained high accuracy of 83.3% consistent with previous results reported in a similar task\textsuperscript{3}. All classification algorithms allowed for higher accuracy than our baseline.

### Table 2. Classification algorithms and accuracy obtained with the held-out testing set.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameter Settings</th>
<th>Feature Context Used (Yes/No)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DummyClassifier (baseline)</td>
<td>strategy=most_frequent</td>
<td>No</td>
<td>46.55%</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>Default</td>
<td>No</td>
<td>78.99%</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>Default</td>
<td>Yes</td>
<td>81.84%</td>
</tr>
<tr>
<td>LinearSVC</td>
<td>$C=0.024$</td>
<td>Yes</td>
<td>84.21%</td>
</tr>
<tr>
<td>DecisionTreeClassifier</td>
<td>Default</td>
<td>No</td>
<td>79.20%</td>
</tr>
<tr>
<td>DecisionTreeClassifier</td>
<td>Default</td>
<td>Yes</td>
<td>78.92%</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>Default</td>
<td>No</td>
<td>79.29%</td>
</tr>
<tr>
<td>RandomForestClassifier</td>
<td>Default</td>
<td>Yes</td>
<td>74.38%</td>
</tr>
<tr>
<td>RandomForestClassifier*</td>
<td>n_estimators=100</td>
<td>Yes</td>
<td>80.29%</td>
</tr>
<tr>
<td>SGDClassifier</td>
<td>Default</td>
<td>No</td>
<td>78.66%</td>
</tr>
<tr>
<td>SGDClassifier</td>
<td>Default</td>
<td>Yes</td>
<td>83.44%</td>
</tr>
<tr>
<td>MultinomialNB</td>
<td>Default</td>
<td>No</td>
<td>75.44%</td>
</tr>
<tr>
<td>MultinomialNB</td>
<td>Default</td>
<td>Yes</td>
<td>75.08%</td>
</tr>
<tr>
<td>MultinomialNB</td>
<td>alpha=0.2</td>
<td>Yes</td>
<td>79.25%</td>
</tr>
<tr>
<td>CRF++</td>
<td>Default</td>
<td>Yes</td>
<td>83.3%</td>
</tr>
</tbody>
</table>

* Top 10000 best features selected with chi-squared in scikit-learn algorithm SelectKBest(chi2,k=10000) from training data.

### Conclusion

This study showed the importance of local contextual features and algorithm parameters fine-tuning when classifying clinical event relations with the clinical document creation time. The highest accuracy was reached by a support vector machine (LinearSVC) algorithm with a grid search for the best $C$ parameter. This accuracy is higher than previously reported for this task\textsuperscript{4}. We found it useful to work with a machine learning framework like scikit-learn that offers a wide selection of classification algorithms. Future work might focus on utilizing these temporal relations to give a summarized view of a patient clinical events timeline.

### References

Introduction
US healthcare systems are rapidly implementing online patient portals linked to electronic health records (EHRs) in order to meet federal Meaningful Use certification criteria. Specifically, Stage 2 Meaningful Use tied financial incentives to portal access by requiring that 50% of patients be offered online access to their health information and 5% of patients use portal websites at least once. Lagging behind other healthcare systems in patient portal implementation, safety net institutions face particularly large obstacles to widespread adoption, including patient populations with limited health literacy and/or English literacy, limited proficiency with digital technology, concerns about breaches of privacy, and mental health and/or substance use conditions. This is the first study to our knowledge that has explored patient portal implementation at the organizational level among safety net healthcare systems striving to meet Meaningful Use measurement thresholds. Drawing on sociotechnical theories in which technology adoption is conceived as a situated, emergent process of organizational change rather than simply a technical matter, we investigated the strategies used to implement patient portals at clinics that successfully achieved Meaningful Use certification. Our goal was to better understand how Meaningful Use criteria, and the portal tool itself, shaped specific implementation activities and outcomes.

Methods
We used a rapid ethnography approach in order to glean insights about a range of contextual factors, practices, and experiences of administrators, clinicians and staff that were relevant to patient portal implementation and efforts to meet Meaningful Use targets. Between August and October 2015, we conducted 5 site visits with community health centers across California. All 5 sites were engaged in Meaningful Use Stage 2 reporting to increase patient registration and use of their newly launched online patient portal websites. We selected the health centers to represent Northern and Southern California, rural and suburban sites, and a range in size (single clinic to a 10+ clinic organization). Sites also represented both high and moderate success in reaching the two Meaningful Use patient-level benchmarks (i.e., 3 had already met the benchmarks and the other 2 were still working on their attestation process). For 4 of the sites, we conducted 1- to 1.5-day site visits that included in-depth interviews with leaders, informal focus groups with frontline staff, observations of clinic practices for patient portal sign-up, and review of marketing materials and clinic workflows. Unable to arrange a site visit at the 5th clinic, we instead conducted in-depth phone interviews with leaders and frontline staff. Interviews and focus groups were audio-recorded and professionally transcribed, and 3 team members wrote detailed fieldnotes. Data analysis followed the constant comparative method, in which themes were identified through repeated readings of the data and group discussions. Interpretations were further informed by sociotechnical theory’s insights about the interrelatedness of social and technical processes in technology adoption.

Results
All five clinics used an array of strategies to ensure that Meaningful Use thresholds were reached. We found that these efforts fell into three broad domains (see Table for additional findings by domain): 1. Aligning interests: health systems strove to ensure that Meaningful Use requirements were prioritized by stakeholder groups responsible for registering patients for portal use, particularly frontline staff. Implementation efforts focused on (a) promoting the portal to staff and clinicians as a strategy for improving patient care and clinic efficiency and (b) ensuring commitment to patient registration as the primary implementation goal. Patient outreach was often limited to basic information about the portal and assistance with the registration process. 2. Articulating existing clinical work with the demands of the tool: health systems mobilized considerable resources to transform existing clinic workflows in order to routinize portal sign-up and patient message routing. Customizing the portal to make it more usable for staff, clinicians and patients was constrained by the inflexibility of the EHR’s functionalities. 3. Defining success: overall the clinics in our study understood successful implementation to be based on (a) how well enrollment procedures and patient message routing were integrated into clinic work and (b) achievement of Meaningful Use targets. For the most part, success was not contingent on use of the portal by patients, or on the impacts of portal use...
on clinic operations and patient health outcomes. In short, clinic priorities were compatible with Meaningful Use mandates but included limited efforts to enhance the relevance and usability of the portal for patients. Sustained use of the portal by a large proportion of patients was a goal deferred to an unspecified future.

### Table of Ethnographic Findings

<table>
<thead>
<tr>
<th>Domain</th>
<th>Theme</th>
<th>Illustrative quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligning interests</td>
<td>Promote the portal primarily to staff and clinicians</td>
<td>“We needed to sell it to providers as less work to track down patients.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“There needs to be an organizational-wide engagement with the portal. We think it’s the right thing to do.”</td>
</tr>
<tr>
<td></td>
<td>Emphasize patient sign-up as means of realizing portal’s promise</td>
<td>“They knew that this was an organizational goal. They knew that they had to get their numbers up.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Their job was to get it to every patient coming through the doors.”</td>
</tr>
<tr>
<td>Articulating work and tool</td>
<td>Adjust clinical work to accommodate portal registration and messaging</td>
<td>“We think it’s the right thing to do, but…we stopped other work to get our enrollment numbers.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“When a site wasn’t meeting the goal, we moved around the staff – like tetris.”</td>
</tr>
<tr>
<td></td>
<td>Difficulty customizing portal to accommodate clinic and patients</td>
<td>“It’s been a struggle…we don’t have access to what the patient sees.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“The portal can work in Spanish, but only the navigation…patients want to act in their own language.”</td>
</tr>
<tr>
<td>Defining success</td>
<td>Reaching Meaningful Use metric</td>
<td>“We were very driven by Meaningful Use…we continue to do a lot of work to make sure that incentive money is coming to us.”</td>
</tr>
<tr>
<td></td>
<td>Lack of emphasis on relevance and usability of portal for patients</td>
<td>“…our managers [say], ‘Don’t market too much. We don’t really want all our patients signed up. We just need to reach the 5% requirement.”’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“…the driver becomes meeting certain Meaningful Use standards sometimes taking priority over meaningful use for our patients.”</td>
</tr>
</tbody>
</table>

### Discussion

We found that safety net health systems engaged in a diverse array of strategies to ensure a good “fit” between Meaningful Use requirements, patient portal functionalities, and clinic work processes. To a large extent, implementation efforts resulted in the successful achievement of Meaningful Use goals and a belief among clinic staff that the patient portal is a means of improving patient care. However, adoption activities primarily served the goal of increasing patient registration rather than enhancing patients’ skills and motivation to use the portal, or ensuring that the tool is useful, relevant and beneficial to patients. Ultimately, Meaningful Use was achieved without addressing significant mismatches between EHR platforms and the priorities and circumstances of safety net patient populations. Future policies to promote patient engagement with EHRs should a) consider the demographics and needs of safety net populations into consideration; b) require strategies to increase patients’ computer literacy and access to internet-connected devices; c) include incentives for more meaningful outreach and collaboration with patients; and d) ensure that patient portal websites are available in languages other than English.

### References

Deep Learning Recommendation of Treatment from Electronic Data

Melissa Aczon, Ph.D., David Ledbetter, Long Van Ho, Alec Gunny, Randall Wetzel, M.D.
Children’s Hospital Los Angeles, Los Angeles, CA

Introduction

Electronic health records (EHR) containing the medical and treatment histories of patients are collected in hospitals nationwide. Advanced information extraction techniques from the machine learning community increasingly have been applied to these data to provide insights and predictions that help physicians provide the best care to their patients.\textsuperscript{1, 2} In this work, we utilize a Recurrent Neural Network (RNN) to learn a model that can predict the effect of treatments on an individual patient’s vitals and probability of survival. The model enables a clinician to insert a therapy and assess whether the vital and mortality predictions are perturbed in a useful fashion. We use mortality prediction for two primary reasons:

- Mortality is a well-defined, short-term truth that lends itself to an objective assessment of a model’s ability to understand the different correlations amongst the many variables that define a patient.
- The more accurate this function is as a predictor for mortality response, then the more reliable it is as an objective function to optimize when making treatment recommendations.

A novel aspect of this approach is its ability to enable treatment recommendation without an explicit diagnostic step. No explicit clusters of similar patients are made to determine a diagnostic or prognostic significance. In effect, however, clustering of similar patients is implicitly accomplished during model training.

Methods

We leveraged EHR of more than 8,000 patients who were admitted to the Pediatric Intensive Care Unit (PICU) of Children’s Hospital Los Angeles (CHLA) over a period of more than ten years. The recorded data include irregular time series of patient vitals (e.g., heart rate), laboratory measurements (e.g., glucose levels), drugs given (e.g., epinephrine), and interventions (e.g., intubation). Each patient’s PICU encounter also has a well-defined outcome: she either lived or died at the end of her stay.

Figure 1 illustrates a high-level overview of our scheme, where the kernel is an RNN utilizing a Long Short-Term Memory (LSTM) architecture.\textsuperscript{3} Inputs to the network are patient vitals and treatments, while outputs are predictions for probability of survival and a subset of the vitals: heart rate, diastolic blood pressure, systolic blood pressure, respiratory rate, and pulse oximetry.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure1.png}
\caption{Overview of RNN framework applied to CHLA PICU Data}
\end{figure}

The RNN approach is inspired by our view of a patient’s trajectory as a continuous dynamical system that consists of many variables internal to the patient (vitals and labs, denoted by the vector $X_V$) interacting with administered treatments (drugs and interventions, denoted by the vector $X_T$): $d/dtX_V(t) = F[X_V(t), X_T(t)]$. Any finite difference scheme that approximates this continuous system takes on the form

$$X_V(t_{n+1}) = G[X_V(t_n), X_T(t_n), X_V(t_{n-1}), X_T(t_{n-1}), \ldots, X_V(t_{n-k}), X_T(t_{n-k})].$$

The RNN kernel is trained to learn the function $G$ which describes the evolution of patient state from internal and treatment history. In the above equation, $t_n$ denotes the current time, and $t_{n+1}$ is the desired time for a future prediction. The difference, $\Delta t_n = t_{n+1} - t_n$, is an input to the kernel at the current time, along with the associated patient physiologic and treatment data at that time: $[X_V(t_n), X_T(t_n)]$. This set-up is amenable to a real-time scenario where
a clinician would have the ability to insert alternative treatments and specify how far ahead into the future the model makes its predictions. To enable quantitative evaluations of the model’s vitals predictions, $\Delta t_n$ was chosen such that $t_n + \Delta t_n$ corresponds to a time at which recorded measurements are available. In the PICU database, $\Delta t_n$ ranges from minutes to several hours.

A single patient can have multiple encounters, where an encounter is defined as a contiguous time between a patient’s admission into and discharge from the PICU. This is an important point for validation because vitals and treatments can be more highly correlated between encounters of the same patient than one would expect between two distinct patients. When the database was partitioned into training and test sets, multiple encounters of a single patient were not allowed to be separated. This partitioning resulted in a training database with 5831 unique patients spanning 7679 encounters and a test database with 2520 unique patients spanning 3311 encounters.

**Results and Discussion**

The RNN’s model predictions were assessed using ROC AUC at clinically significant hours: [1, 3, 6, 12]. In other words, after observing the patient’s state for $n$ hours, the RNN would predict whether the patient survived or died at the end of their stay. Figure 2a illustrates the model’s performance in mortality predictions on the test set. As the observation window increases there is a corresponding increase in accuracy of the mortality predictions evident in the ROC AUC.

Vital predictions are made for the subsequent timestep which range from a minute to several hours into the future. The mean absolute error over the entire test set is 1.5%. For comparison, a simple forward fill, which can be regarded as a baseline, yields an error of 2.2%. The RNN model reduced the error by 30%. Figure 2b displays predictions for an individual patient (from the holdout set) who died in the PICU. In this particular case, a catastrophic deterioration, from which the patient never recovers, can be seen two days before death.

![ROC Curve](image)

Figure 2: (a) Mortality assessment via Receiver Operating Characteristic curves. (b) Example of vitals and mortality predictions through time for an individual patient.

**Summary**

We utilized recurrent neural networks to generate predictions for vitals and mortality. The RNN-based framework is amenable to real-time scenario simulation: it enables the effects of treatment decisions to be modeled prior to execution. The goal of this work is to build a clinical decision support tool that evaluates the impact of alternative treatments and displays to the clinician those treatments that have the highest probability of positive impact to the patient.

**References**

Using data assimilation to forecast post-meal glucose for patients with type 2 diabetes

David Albers1, PhD, Matthew Levine1, BA, Andrew Stuart2, PhD, George Hripcsak1, MD MS, and Lena Mamykina1, PhD

1Department Biomedical Informatics, Columbia University, New York, NY
2Department Computing and Mathematical Sciences, Caltech, Pasadena, CA

Introduction

Type 2 diabetes (T2D) is a disease with dire consequences if not treated to minimize its effects, is costly ($245 billion a year in 2012), and is prevalent, affecting over 8% of the US population [1]. Treatment of T2D relies primarily on self-management of nutrition, exercise, and medication and is highly dependent on the individual. While there exists general guidelines for healthy eating, each individual with diabetes must find their own way to translate these guidelines into daily meal choices [2]. This is usually accomplished through a lengthy, often complex and opaque, trial and error process in which individuals experiment with different foods and observe their impact on blood glucose levels. As is the case with many complex nonlinear processes, humans are not well suited to finding optimal solutions that are context-specific and personalized without machine-based support. The methodology we use here, data assimilation (DA) [3], is data-science technique used in a wide range of applications, e.g., weather forecasting or aerospace engineering. Data assimilation marries data with human knowledge encoded in mechanistic models, personalizes model parameters, and makes a real-time forecast based on the model and the data. In applications where data are sparse and the system is nonstationary, data assimilation overcomes data-quality problems by leveraging knowledge encoded in mechanistic models. DA, through the continuous personalization of mechanistic glucose/insulin models, computes nutrition and glucose-based post-meal glucose forecasts, creating a quantitative, real-time link between nutrition choice and blood glucose.

Methods

We prospectively collected at least 21 days of nutrition and blood glucose data from patients in a T2D intervention study. The data were collected using the Mealzer mobile application and constrained to realistic T2D outpatient data collecting: sparse finger prick glucose measurements before and after meals and pictures of meals whose nutrition content was estimated by trained nutritionists. We then selected two mechanistic endocrine models, a relatively simple ultradian model [4] and a complex meal model [5] with detailed modeling of metabolism. Both models can take blood glucose, insulin, and carbohydrates as input variables, have between 6 and 12 state variables including plasma glucose and insulin and have between 30 and 70 other physiologic parameters. We use several DAs to integrate data with models, including a dual unscented Kalman filter (UKF) [6], particle filters, and other inverse-problem Bayesian methods [3]. Using the patient data and the DA, we generate personalized, real-time nutrition-based post-meal glucose forecast for both models. To evaluate which model performed better, we use both mean squared error (MSE) of the model forecast compared to training set and improved with more accurate nutrition estimates, but forecasting a given post-meal glucose value was relatively insensitive to nutrition estimate accuracy. Finally, careful off-line parameter estimation yielded more accurate forecasts.

Results

Within 20-30 meals both models personalized and converged to the individual patients, and both models forecasted post-meal glucose well, cf. Fig. 1. The ultradian model minimized the MSE while the meal model minimized the KL divergence between kernel density estimate of the measured and forecasted glucose densities. The disagreement between these model evaluation metrics has a clinical implication: model evaluation metrics must be chosen to maximize accuracy relative to a clinical goal, such as resolving mean glucose. We also found the models to be sensitive to nutrition estimate: personalizing the model parameters required a consistent training set and improved with more accurate nutrition estimates, but forecasting a given post-meal glucose value was relatively insensitive to nutrition estimate accuracy. Finally, careful off-line parameter estimation yielded more accurate forecasts.

Conclusion

Data assimilation methods, even with limited training, can provide accurate, personalized, nutrition-based glucose forecasts in real time. While both models personalized to the individual, they provide different “optimal” knowledge; further research is needed to determine which knowledge type is more relevant to promote individual decision-making in diabetes. Given the number of mechanistic glucose models available [8], we must evaluate which model will be of greatest potential impact, subject to an evaluation metric chosen to yield the best self-care-based outcome. Model parameter estimation --- personalization and adaptation of the model --- is best done using consistently sourced and accurate nutrition information to put the DA into the right parameter region, but the real-time nutrition estimate used to forecast a post-meal glucose can have substantial error and remain accurate. Because of these conclusions, we are hopeful that DA can be used to personalize nutrition-based, post-meal glucose forecasts, allowing individuals with T2D to make informed nutrition choices based on real-time glucose forecasts.

Acknowledgements We acknowledge NLM grant R01 LM06910, NIDDK grant 5R01DK090372, and the Robert Wood Johnson Foundation.
References


Figure 1: The blood glucose measurements and the data assimilation based real-time continuous glucose and on-measurement blood glucose forecasts for days 1-5 and 20-25 for the ultradian model (top) and the meal model (bottom). As more measurements are accumulated, the data assimilator adjusts the parameters given the individual’s data allowing for more accurate forecasts as more measurements are accumulated as can be seen by comparing the left and right plots.
A Terminology Framework for Nursing Documentation Redesign

Deborah Ariosto, PhD, MSN, RN
Vanderbilt University Medical Center, Nashville, TN

Abstract

**Background:** In response to nurse staffing ratio changes, an initiative to decrease time spent documenting was launched. **Methodology:** Informaticists and clinical staff redesigned documentation using the Clinical Care Classification System\(^1\) for a problem based, modified charting by exception approach. **Results:** A 30% decrease in volume charted contributed to successful adoption and high nurse satisfaction. **Discussion:** Staff nurse involvement guided by a terminology framework reflective of professional practice model\(^2\) shaped a better product than either could alone.

Introduction

Nurses are the primary caregivers in inpatient settings. Studies report that much more time is spent documenting than delivering care\(^3\). In response to staffing ratio changes, at a large southeastern academic medical center, a team of clinicians and industrial engineers were deployed to observe and interview nursing staff to identify opportunities to save nursing staff time. Top priorities identified by staff nurses were unnecessary/redundant documentation, finding equipment, and better communication with the care team. In 2014, executive leadership endorsed the formation of cross enterprise clinical workgroups to address these priorities. This presentation focuses on the methodology applied to a highly successful documentation reduction initiative completed and rolled out in 2015 across both the adult and pediatric populations.

Methodology

Over fifty clinical staff from acute and critical care in both the adult and pediatric hospitals was challenged to standardize and simplify nursing documentation. Exclusions included Obstetrics, Neonates, and Psychiatry. A representative core group of nurses met weekly to shape a common documentation model using a terminology framework that reflected professional nursing practice endorsed by the American Nurses Association\(^1\).

The Clinical Care Classification System (CCC) was used to organize nursing care concepts and compress into 15 Care Categories to create a high level framework for nursing centric problems (e.g. acute pain, confusion, injury risk), interventions, and goal setting. SNOMED CT\(^4\) concept matching was used to standardize symptoms and intervention detail. Free text narrative statements captured the overall patient response to care and communication/event notes.

Other strategies included a modified charting by exception approach and elimination of most copy/paste functions. Assessments auto-populated the plan of care and easily visualized the daily progress toward problem resolution. Documentation focused diagnosing problems, implementing and evaluating patient response to the plan of care. Documentation that did not reflect this professional practice model or support high value decision support was discouraged.
Results

Several pre/post evaluation approaches were utilized. Timed classroom simulations by clinical staff showed on average, a 25 minute reduction in time spent documenting. Real-time documentation of nursing care on pilot units also showed on average, 25 fewer minutes spent documenting. In addition, the overall volume of data documented per patient per day decreased by 30% overall when comparing January 2015 (pre-implementation) to January 2016 (post-implementation). The graph below shows the relative changes in documentation by adult and pediatric populations.

![Change in Nursing Documentation](image)

**Change in Nursing Documentation**

<table>
<thead>
<tr>
<th>Fields/Patient/Day</th>
<th>JAN 2015 (PRE)</th>
<th>JAN 2016 (POST)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADULT - ALL</strong></td>
<td>354</td>
<td>246</td>
</tr>
<tr>
<td><strong>ADULT - ICU</strong></td>
<td>790</td>
<td>583</td>
</tr>
<tr>
<td><strong>ADULT - ACUTE</strong></td>
<td>254</td>
<td>167</td>
</tr>
<tr>
<td><strong>ADULT - ED</strong></td>
<td>146</td>
<td>114</td>
</tr>
<tr>
<td><strong>PEDS - ALL</strong></td>
<td>337</td>
<td>249</td>
</tr>
<tr>
<td><strong>PEDS - ICU</strong></td>
<td>991</td>
<td>776</td>
</tr>
<tr>
<td><strong>PEDS - ACUTE</strong></td>
<td>221</td>
<td>152</td>
</tr>
<tr>
<td><strong>PEDS - ED</strong></td>
<td>79</td>
<td>72</td>
</tr>
</tbody>
</table>

Discussion

Staff has had an overwhelmingly positive response in that they have more time to spend with their patients and take needed breaks. Rounding during implementation reassured leaders of the value of their investment in staff time and resources and gave staff an opportunity to show their work and tell their stories. One poignant story shared was: “I can take my toddler off the ventilator and sit with him, and let him be a child again.”

Providing nurses with a nursing terminology framework that they could shape and own was key. While the informaticists knows the value of standards for data re-use, interoperability, decision support, and system stability, the clinician knows their own practice and unit needs. Working together, documentation re-design achieved both goals.

References

1. [http://www.sabacare.com](http://www.sabacare.com)
3. Hendrich A, Chow M, Skierczynski, Zhenqang L. A 36-Hospital Time and Motion Study: How Do Medical-Surgical Nurses Spend Their Time? The Permanente Journal/ Summer 2008/ Volume 12No. 3
Clinical Decision Support for Worker Health: A Qualitative Study of Five Primary Care Settings

Joan S. Ash, Ph.D.¹, Dian Chase, Ph.D., R.N., F.N.P.¹, Sherry Baron, M.D., M.P.H.², Margaret S. Filios, M.Sc, R.N.³, Richard N. Shiffman, M.D., M.C.I.S.⁴, Stacey Marovich, M.S., M.H.I.³, Jane Wiesen, Ph.D.¹, Genevieve B. Luenksman, Ph.D.³

¹Oregon Health & Science University, Portland, OR, USA
²City University of New York, New York, NY, USA
³National Institute for Occupational Safety and Health/CDC, Cincinnati, OH and Morgantown, WV
⁴Yale University, New Haven, CT, USA

Introduction
Although primary care providers are usually the first to see patients with medical issues such as asthma symptoms that may be caused by workplace exposures, they do not routinely ask patients about their work ¹⁻⁵. Occupational health physicians have developed evidence-based guidelines for helping to manage many such patients, but primary care providers are rarely aware of their existence⁶. Though computerized clinical decision support (CDS) holds potential for increasing awareness and providing guidance, it must be developed and implemented to fit the context and workflow of those who would benefit from having the information⁷.

This qualitative evaluation project was part of a larger project funded by the National Institute for Occupational Safety and Health (NIOSH) of the Centers for Disease Control and Prevention. The NIOSH project is designing, developing, and piloting CDS for the health of working patients in primary care outpatient settings. The first step in this project was the development of three knowledge resources (KRs) containing evidence-based information, decision logic, scenarios and examples of use. The KRs were prepared by three subject matter expert (SME) groups for three topics that are related to the health of patients who work and considered especially pertinent to a primary care practice. The three KRs focused on dealing with work environment factors that impact the management of a chronic disease (diabetes), guidance for return-to-work after lower back pain diagnosis not related to work, and diagnosis and management of work-related/work-exacerbated asthma.

The goal of this qualitative study is to identify the barriers and facilitators related to CDS for the clinical management of working patients in a variety of primary care settings, including assessment of the technical and organizational feasibility of implementing the CDS represented by each KR.

Methods
We used the Rapid Assessment Process (RAP), as previously described⁸, for studying five organizations, though we adapted it to include graphical elicitation interviews⁹ to evaluate the KRs. Institutional review boards (IRBs) at NIOSH and OHSU and each clinical site with an IRB approved or reviewed the study. Sites were selected for diversity in size, geography, organization, and electronic health record (EHR) product used. We selected subjects based on their roles. Data collection consisted of semi-structured and graphical elicitation interviews⁹ and observations in clinics. Broad areas for exploration included 1) present CDS, 2) clinic and individual workflow, 3) information already available to help manage care of working patients, such as the patient’s work information, 4) barriers and facilitators of CDS for the clinical management of working patients, and 5) evaluation of the three KRs. Both template and grounded hermeneutic methods were used for analyzing data¹⁰⁻¹¹.

Results
We conducted five site visits between July 21st and September 11th, 2015. We interviewed 41 clinicians not deeply involved in informatics, 23 individuals who were informaticians or information technology specialists, 15 managers or staff members, and four quality improvement specialists, for a total of 83 interviews. We spent 30 hours observing in ten clinics. Template analysis of data pertaining to the three KRs gathered during graphic elicitation interviews yielded valuable content and context-related information for developers and implementers. Figure 1 includes the themes and subthemes that emanated from the hermeneutic analysis, grouped as facilitators and barriers. The main themes are of equal importance for acceptance of the proposed CDS, while the subthemes vary in strength of importance and specific relevance to worker health CDS.
Figure 1. Facilitators and barriers to worker health CDS

Discussion and Conclusion
Using a qualitative approach which adapted RAP to evaluate three knowledge resources related to clinical management of working patients, we identified themes representing both barriers and facilitators that can be considered as the CDS is further developed and implemented. While some of these factors are similar to those discovered in prior CDS research, others, such as changing work patterns and roles and the sensitivities and complexities of work-related information, are novel and deserve further investigation.

Acknowledgements
This project was supported by CDC/NIOSH Contract 200-2015-61837 as part of NORA project #927ZLDN. We appreciate the help of the following at our sites: Rose H. Goldman, MD, MPH, Laura Brightman, MD, Stacey Curry, MPH, Larry J. Knight, MSA, MSHRM, Deborah Lerner, MD, Herb K. Schultz, Michael Rabovsky, MD, Joseph Conigliaro, MD, MPH, and Nicole Donoghue, MPH.

References
Translating Hereditary and Familial Cancer Guidelines into Clinical Decision Support

Ayesha Aziz, PhD,1 Lance Pflieger,2 Bryce Carter,2 Joshua Schiffman, MD,2 Brandon M. Welch, MS, PhD1

1Medical University of South Carolina, Charleston, SC, 2University of Utah, Salt Lake City, UT

Abstract:
Family Health History (FHx) is one of the most important cancer risk factor. To develop effective evidence-based CDS tools for hereditary and familial cancer, there is a need to translate knowledge from cancer guidelines into computable CDS knowledgebase. We conducted a systematic analysis of available familial and hereditary cancer guidelines to identify gaps in published guidelines, and challenges in informatics standards and terminologies that would impede the ability to implement these guidelines as standard-based CDS.

Introduction
A family history (FHx) of cancer is a major contributor to one’s risk for cancer. For a clinician, a complete and accurate FHx helps to identify a patient’s cancer risk and initiate risk-reducing strategies at an earlier stage. A number of recognized medical and government bodies (e.g. U.S. Preventive Services Task Force, National Comprehensive Cancer Network, American Cancer Society) have published cancer care and prevention guidelines that incorporate family history of cancer. While collection of family health history from patients has traditionally been sub-optimal,1 current efforts in health IT are expanding the ability to collect a patient’s structured family health history. Notably, HL7 developed and published a pedigree model with the aim of representing FHx in a detailed format.2 Additionally, Meaningful Use Stage II guidelines include FHx as an optional menu set item for electronic health record certification.3 As a result, EHRs, researchers, and vendors are also developing informatics tools to improve the collection of a structured FHx.4 With a patient’s FHx information more readily available, the opportunity exists to translate established cancer guidelines related to family cancer into computable clinical decision support and implemented within FHx tools. While there have been previous efforts to implement family cancer guidelines in CDS, these efforts were built upon proprietary systems and limited to a single guideline and a single cancer use case (e.g. breast cancer).5,6 Moreover, these implementations lacked compliance with the current health IT standards and terminologies, and provided no mechanism of updating the CDS knowledgebase as improvements in the published guidelines became available. Therefore, given the opportunity to deploy cancer guidelines on new, state-of-the-art FHx collection approaches, we conducted a systematic analysis of cancer guidelines that could use FHx information. We sought to (1) assess the feasibility of converting currently available guidelines into standards-based CDS, (2) identify gaps in informatics standards and terminologies, and (3) highlight challenges within current guidelines that limit their ability to be implemented. To our knowledge, this is the first systematic analysis of hereditary and familial cancer guidelines that assess ability to be implemented in standards-based CDS.

Methods
We identified a set of 20 professional, advocacy, medical, and government organizations that publish guidelines for eight cancers with common heritable and familial risks. Using pre-specified search terms, we identified logical (IF.. THEN..) statements related to cancer in the family. Independent reviewers (AA and LP) who are specialized in Biomedical informatics and computerized clinical decision support conducted the review. For each statement, content analysis was conducted to determine if the statement could be represented using current Meaningful Use stage 2 FHx health IT standards, which include the full HL7v3 pedigree model, HL7v3 pedigree model limited (first degree relatives only), SNOMED full, SNOMED limited (first degree relatives only). We identified whether or not an existing statement can be recorded as structured data for each standard. Assessments of each standard were reported as either (1) explicit or derivable, (2) conditional to fuzzy terms in the guideline, (3) not explicit or derivable, (4) conditional to fuzzy terms and not explicit, or (5) uninformative statements. After each independent reviewer completed their initial analysis, assessments were compared. Disagreements in assignment were discussed and consensus was built by the reviewers with assistance from an additional reviewer/mediator (BW).
Results
We identified a total of 78 guidelines that included statements or recommendations meeting our inclusion criteria. A total of 589 statements were selected for analysis. Out of these statements, Breast, colorectal and ovarian cancer represented roughly half of the total number of statements studied. The analysis revealed that only about 60% of statements derived from the selected guidelines could be implemented using HL7 pedigree model and less than 50% could be implemented using current meaningful use stage 2 standards. The discrepancies between guidelines and standards result from (1) gaps in data model or terminologies, and/or (2) inadequate or poorly represented guidelines (e.g. inadequate or ambiguous definitions). A more detailed description of these discrepancies will be described during the podium presentation.

Discussion
This systematic analysis identified significant gaps that would impede the ability to implement these guidelines as standard-based CDS. Some of the issues could be resolved if the guidelines included more clear or explicit definitions. Other issues could be resolved if the data models or terminologies were improved. The guidelines could be improved if the developers explicitly defining terms such that they can be properly implemented. For example, in several guidelines the phrase ‘strong family history of {disease}’ was used with no clear definition as to which relatives are included or what qualifies as ‘strong’. This issue is not limited to cancer family history guidelines, rather is a common issue faced by CDS knowledge engineers.7 Gaps in implementation vary by the health IT standards used. The full HL7 pedigree model provides the most complete coverage, whereas the minimum requirements using SNOMED CT and first degree relatives as set forth by Meaningful Use Stage II, provides the least guideline coverage. Nevertheless, most guidelines could be implemented if definitions are created and/or data models and terminologies are extended. Therefore, results from this work can be used to inform guideline developing organizations and standards organizations. The systematic method for assessing cancer guidelines could be utilized for other clinical domains.

References

Acknowledgements: This work was partially supported by T15 LM-007124
Evolution of an integrated Patient & Population Analytics Portal (iP3)

University of Illinois at Chicago, IL 60612

Abstract
Effective care coordination especially when managing complex populations requires partnerships of multidisciplinary teams, physicians, specialists, pharmacists, care organizations, payers and institutions. Another requirement for a successful model is to have effective data integration and analytics platform. In this presentation, we intend to share our experiences and evolution of an integrated patient analytics portal for management of The University of Illinois CHECK (Coordination of Healthcare for Complex Kids) project, which aims to create and test a “medical neighborhood” comprised of over 40 practices in Cook County. The program intends to target 6,000 children with complex conditions. We established a centralized data warehouse to accumulate and process clinical, claims, assessment, engagement and school data that provide systemic insights into the program and aid in meeting program’s aims while managing risks and overcoming significant people, process and technology related challenges. We present our methodology of mapping program goals to analytics goals that eventually led to creation of the portal iP3, an Integrated Patient & Population Analytics Portal that can be generalized for future healthcare research studies.

Introduction
CHECK (Coordination of Health Care for Complex Kids) is designed as a model that improves care for populations with specialized needs, with a focus on the specific priority area of high-cost pediatric and young adult populations. Specifically CHECK aims to: 1) to reduce both inpatient admissions and emergency room visits, 2) to reduce missed school days and, 3) to increase patient satisfaction, engagement that will eventually reduce healthcare costs. To accomplish these goals, UI Health partnered with multiple community primary health care providers, community mental health providers, community dental health providers and social services providers. Data from over 10 disparate sources needed to be collected and transformed at UIC’s centralized warehouse to a) Perform ETL on Medicaid Claims from multiple care organizations, Clinical EMR data from multiple provider organizations, Assessment data from RedCap and Third Party tools, Engagement Touches from Care Coordinators, Patient Communication logs and Public School data, b) Develop Risk/Eligibility Criteria to identify eligible patient cohorts, c) Track ongoing enrollment and disenrollment process, d) Facilitate data exchange to enable assessments performed by a third party, e) Monitor the financial impact of the program continuously in form of PMPM and Service Utilizations across targeted diseases and encounter types, f) Create integrated patient analytics portal for program leadership to facilitate monitoring and decision-making based on financial and clinical indices and g) Identify Bright Spots where CHECK is making the greatest impact.

Methodology
To meet program goals and design iP3, our informatics team worked closely with various stakeholders from executive management, community services management, clinical leadership, mental health and dental professionals, third party outreach & assessment centers and Centers for Medicare & Medicaid Services (CMS). Each stakeholder viewed the program and its aims through the prism of their domain that added complexity to defining metrics, informatics goals that will eventually lead to reports and visualizations for consumption by different audiences. We established systematic processes to engage multiple stakeholders to avoid duplication of efforts and bring everyone to a common platform. We mapped the program aims to measurable outcomes and identified the required data elements. From identifying data sources for required data elements to establishing ETL processes with numerous MCOs to developing coherent view of the patient across all MCOs, the whole process was executed by breaking down program’s functions into separate groups that eventually led to nominating functional owners of both Clinical and Informatics domains.

Result and Discussion
In this presentation, we glanced over our methodology to design an integrated Patient & Population Analytics Portal (iP3) that involved handling vast amount of heterogeneous disparate data. As of today, iP3
allows CHECK leadership to perform these activities in real-time - Identify potential cohort for the program based on one of many available Risk Stratification algorithm such as NYU ED & UCSD CDPS, create population profiles based on Medicaid Costs distributions such as PMPM (Per Member Per Month) and service utilization per 1000 across various primary diseases, time series analysis of engaged population costs. iP3 used geo-mapping techniques to visualize CHECK population across various stages of engagement, diseases, risk score across various zip codes within Chicago areas to identify potential disease concentration based on environmental factors of the neighborhood and to understand healthcare accessibility based on the potential distance between the patient and the care. Next logical step for iP3 would be doing predictive analytics to identify high-risk patients early on during the patient engagement stage and develop real-time mechanism for community health worker to get up to the minute information of unexpected hospitalizations directly from the hospital of their assigned patients. iP3 is at an exciting place where opportunities are limitless but to navigate substantial risks and challenges successfully we will continue improving our methodology and processes to materialize those opportunities.

THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ILLINOIS
Project Title: "University of Illinois CHECK (Coordination of Healthcare for Complex Kids)"
Geographic Reach: Illinois
Estimated Funding Amount: $19,581,403
Show-Me ECHO: ENDO ECHO – Changing the landscape of endocrinology and diabetes care in Missouri

Mirna Becevic, PhD, MH1,2, Uzma Khan, MD3, Omer Malik, BA2,4, Rachel Mutrux, BA1; Suzanne Boren PhD, MH4, Sarjukumar Panchal, PharmD5; Arpit Aggarwal, MD6, Cameron Lindsey, PharmD, BC-ADM, CDE, BCAC7, Carolyn Edison, ADM-NP, Ed.D5, Karen Edison, MD1,2

1University of Missouri, Department of Dermatology, Columbia, MO, 2University of Missouri, Missouri Telehealth Network, Columbia, MO, 3University of Missouri, Department of Medicine, Columbia, MO, 4University of Missouri, Department of Health Management and Informatics, Columbia, MO, 5University of Missouri, School of Medicine, Columbia, MO, 6University of Missouri, Department of Psychiatry, Columbia, MO, 7University of Missouri-Kansas City School of Pharmacy, Kansas City, MO

Introduction

Background Extension for Community Health Care Outcomes (ECHO) Project is a novel and groundbreaking way of utilizing telemedicine technologies to deliver care and education to rural and underserved areas1. By definition, Project ECHO strives to reduce health care disparities between urban and rural regions by providing evidence-based and state-of-the-art medical education and mentoring to rural providers. Show-Me ECHO Project is a replication of the Project ECHO introduced by Dr. Sanjeev Arora in 2003 at the University of New Mexico2. The Show-Me ECHO project was initiated in November of 2014 with two pilots – autism and chronic pain management. Following the state appropriation of $1.5 million, four other ECHO projects started in July of 2015: endocrinology, dermatology, hepatitis C, and pediatric asthma. Show-Me ECHO is a partnership between an academic medical center – University of Missouri (MU), Missouri Primary Care Association (MPCA), and the State of Missouri, formed to address the lack of specialty care in rural Missouri.

Objective We sought to examine how the ECHO project can potentially influence changes in self-efficacy among primary care providers (PCPs). Provider self-efficacy may be linked to both provider and patient behavior change, patient compliance, and patient outcomes

Methods

Design This was a cross-sectional study of participating providers, which was a part of the Show-Me ECHO required program evaluation. It was a pre-post study that included the same participants, as the repeated cross-sectional surveys provide useful information regarding trends over a period of time3. Participating providers were asked to complete a 14-question self-efficacy survey prior to attending a Show-Me ENDO ECHO first block, and again after their participation. The survey was divided into three sections: 1) confidence in screening for diabetes, 2) confidence in medication and treatment for diabetes, and 3) confidence in patient education.

Setting Program evaluation took place at the University of Missouri from September 2015 – November 2015. The hub team of specialists was located at the University of Missouri, while participating providers at spoke sites were from different regions of the state (Picture 1).

Interventions Weekly 1.5 hour sessions were offered via telehealth to participating providers across the state. Each session was made up of two parts – a case presentation by a participating primary care provider, and a continuing medical education (CME) - approved didactic presentation. Case presentations were done in a de-identified format. The didactic portion of the fall 2015 sessions consisted of the initial “Introduction to ENDO ECHO” session, followed by three blocks: a) thyroid block, b) diabetes block 1, and c) diabetes block 2.

Main outcomes measurement Since the Show-Me ECHO project is in its infancy, at this stage of program evaluation we are not able to analyze provider and patient behavior changes and full impact of ENDO ECHO on Missouri communities. In partnership with the MU Office of Social and Economic Data Analysis (OSEDA), we will analyze the MO HealthNet (Missouri Medicaid) claims data: PCP prescription patterns, emergency department visit costs, diagnosis codes, patient geographic region, HEDIS codes, and outpatient visit costs. Self-efficacy surveys were collected and managed using the REDCap electronic data capture tool hosted by School of Medicine at MU4. Tableau was used to analyze the data5.
Results:
During the Fall 2015 ENDO ECHO block there was a total of 12 sessions with 12 CME-approved didactics. ENDO ECHO was attended by 34 unique attendees, of which 9 active case presenters presented 34 cases. Self-efficacy survey was completed by 44%. The small response rate may be attributed to the fact that participating providers did not receive personalized invitations to participate, and that there were no incentives provided for participation, which are known strategies used to increase rather low healthcare research responses. Overall, self-efficacy increased for all participants. However, in some individual cases it decreased for participating PCPs. The authors associate the decrease of self-efficacy with overestimating one’s knowledge prior to attending educational sessions and subsequently realizing gaps in knowledge. Figure 1 indicates participating PCPs’ reasons for attending ENDO ECHO sessions. The highest increase in overall self-efficacy was in initiating and prescribing insulin injections to patients (pre self-efficacy score was 68.40, post 85.00). The smallest overall increase was in advising self-monitoring of blood glucose (from 78.80 to 82.25). Individual participants indicated decrease in self-efficacy in identification of patients that need screening for diabetes (from 100 to 90), screening patients for neuropathy (from 74 to 56), and cardiac problems (from 70 to 50).

![Image](picture1.png)  
**Picture 1: ECHO ENDO Session**

![Image](figure1.png)  
**Figure 1: Reasons for joining ENDO ECHO**

Discussion:
ENDO ECHO uses disruptive innovation technologies to deliver specialty care and education to rural communities. Disruptive innovation technologies utilize existing technologies, in this instance telehealth, to enhance patient care, PCP’s education, and access to care with minimal or no additional cost⁶. Continuous program evaluation is necessary to ensure the success of the project. Provider self-efficacy surveys were instrumental in providing the specialty hub team with information necessary to create a successful didactic curriculum. In other words, topics with either increase or decrease in self-efficacy were reviewed and additional lectures on those topics were provided, as they were indication that more knowledge in those areas might be needed. MO HealthNet (Missouri Medicaid) claims data will be utilized to tailor the didactic sessions to the needs of the providers, but also to target recruitment of Missouri providers for ENDO ECHO and Show-me ECHO.

References
Little Impact of Highly Specific Alerts for Potassium-Increasing Drug-Drug Interactions — A Randomized Controlled Trial

Jürg Blaser, PhD1, Patrick E. Beeler, MD1, Markus Schneemann, MD2, Emmanuel Eschmann, MD1

1Research Centre for Medical Informatics, University Hospital, Zurich, Switzerland; 2Division of Internal Medicine, University Hospital, Zurich, Switzerland

Introduction

Drug-drug interaction (DDI) alerts are often overridden due to their low specificity. Approaches addressing this issue include (a) focusing on high-priority DDIs, (b) enhancing alerts by displaying critical laboratory data, (c) considering patient data in order to suppress unspecific alerts, (d) timing of alerts (i.e. when an adverse reaction becomes evident or when a DDI is insufficiently monitored, as opposed to systematic alerts at the point of medication order entry), and (e) removing alerts when the triggering conditions are no longer met. However, most DDI alerts focus on a selection of only a few of these approaches and clinical studies evaluating their impact revealed poor adherence and low appropriateness of overrides.

Therefore, we combined all approaches and applied the resulting concept to potassium (K⁺)-increasing DDIs, the most important high-priority DDI at our hospital. The purpose of the study was to determine the effect of this enhanced concept on both process parameters and clinical outcome.

Methods

The study was a cluster randomized controlled trial. All inpatient clinics of the University Hospital Zurich, a teaching hospital in Switzerland with 37,000 inpatient stays per year, were assigned to the intervention or control group. We used replacement randomization, controlling for the frequency of K⁺-increasing DDIs and insufficient K⁺ monitoring. The intervention period lasted one year, starting from 1/7/2014. Data collected during stays in intensive care units were excluded since electronic order entry is not provided in these units.

Both, at onset of and during K⁺-increasing DDIs, the physicians of the intervention group were (i) reminded to monitor the serum K⁺ (if current K⁺ unknown and no K⁺ level measurement ordered), (ii) alerted of the risk for hyperkalemia (most recent serum K⁺ ≥4.9mEq/l and <5.5mEq/l), and (iii) alerted of hyperkalemia (serum K⁺ ≥5.5mEq/l). Alerts and reminders were only displayed as long as the triggering conditions were met. The alerts displayed the prescribed K⁺-increasing drugs, the most recent serum K⁺ level, the glomerular filtration rate and included detailed advice. All reminders and alerts were non-interruptive. They were calculated and logged for all patients, but were only displayed in the electronic health records of the intervention group.

The predefined primary endpoint for assessing the impact of the intervention was the frequency of prolonged K⁺-monitoring intervals during DDI-periods, defined as >72h-periods without K⁺-monitoring. Secondary endpoints of the study included the occurrences of hyperkalemia or deaths during DDIs and the median length of the K⁺-monitoring intervals during DDIs.

Wilcoxon rank sum tests, chi-square tests, and Fisher’s exact tests were used for statistics; p values of ≤0.05 were considered statistically significant.

Results

We analyzed data of 15,272 inpatients in the intervention group and of 18,981 in the control group, including a total of 922,121 drug orders resulting in 2,057 and 2,804 K⁺-increasing DDIs in the intervention and control group, respectively. In the intervention group 869 reminders to monitor K⁺ (1 per 3.2 K⁺-increasing DDIs), 356 alerts of the risk for hyperkalemia (1 per 7.9 DDIs), and 62 hyperkalemia alerts (1 per 45 DDIs) were displayed.

No benefit could be seen with respect to both clinical endpoints: Hyperkalemia occurred during 50 of 2,196 K⁺-increasing DDIs (2.28%) in the intervention group compared to 27 of 1,610 (1.68%) in the control group. The death rate observed in the intervention group was 1.17% (22/1,876) compared to 0.96% (14/1,465) in the control group. No patient died during or within 48 h of a hyperkalemia following a therapy with a K⁺-increasing DDI.
The primary endpoint analysis showed that the frequency of insufficiently monitored DDI-periods was not lower in the intervention group (451 of 9,686 periods >72h; 4.66%) compared to the control group (246 of 6140; 4.06%). In contrast, a statistically significant difference was observed for one secondary endpoint: The length of the K+-monitoring intervals during DDIs was slightly shorter in the intervention group (median 20.3 h) compared to the control group (22.1 h; p<0.001).

**Discussion**

We implemented a highly specific alert concept in order to improve patient safety during K+-increasing DDIs while minimizing the risk of overriding and alert fatigue. However, the randomized controlled trial showed no impact on poorly monitored DDIs and no reduction of hyperkalemia events or death rates. The modest reduction of the median length of the K+-monitoring intervals during DDIs bears little clinical significance.

Not much evidence has been gathered so far for documenting significant clinical benefit of decision support interventions warning for potentially harmful DDIs at the time of order entry. In our study, however, even a sophisticated clinical decision support effort was not successful to improve patient safety regarding the most critical high-priority DDI at our institution.
Linking patients with non-PHI data

Luca Bonomi¹, Xiaoqian Jiang¹
¹ University of California San Diego, La Jolla, CA 92093

Abstract

Patient linkage is important for medical data integration and analysis. Traditionally, this critical step is conducted using protected health information (PHI). We demonstrated that it is possible to link patients with non-PHI data. Our novel model for temporal pattern linkage offers an alternative solution when PHIs are not available and can also serve as a powerful complement to existing PHI-based solutions to break the ties (when limited PHIs lead to ambiguities).

1 Introduction

The medical information systems have undergone significant improvement making it possible to collect various kinds of healthcare data. However, the healthcare data are often fragmented, making the task of reconstructing patient information very challenging. In this work, we proposed a novel method for patients linkage to support Electronic Health Records (EHR) data integration from different sources. Our model takes advantage of the temporal pattern in patients’ non-PHI data (e.g. diagnoses) to link the patient’s complete diagnostic pathway across multiple institutions.

2 Related work

A first step toward the data integration of EHR records consists of extracting useful knowledge form the longitudinal patients data. The authors in⁴ proposed several data mining techniques to predict relationships across diseases. In the traditional database community, a variety works have been proposed to linkage temporal records. Typically, the similarity between two records is computed based on attribute value similarities. However, if entities evolve over time, attributes that subject to change become less reliable indicators for record matching. Thus a temporal model is often constructed to learn how entities evolve over time which determines the weight of each attribute in the matching step. However, these solutions are not very robust to noise and missing data making them not suitable in our setting.

3 Methodology

Our goal is to enable patients linkage using non-PHI data across EHR databases. However, these data (e.g. diagnoses, lab test results) change over time, may be subject to noise, and in general are very sparse, thus, classic record linkage solutions may not be effective. Therefore, we propose a novel technique for linking patients using the similarity between their longitudinal data. Specifically, we model the patients longitudinal data, in our case diagnoses, into sequence, named patients diagnostic pathways. In our approach, we assume that a set $T$, containing the complete diagnostic pathways for some known patients, is given in input. Then, we use this training set to learn the parameters of our similarity model which are used to link a set of new patients (i.e. test set). For each patient in the test set, we randomly divide his/her diagnostic pathway $Z = X \odot Y$ and create two datasets $X \in D_1$ and $Y \in D_2$. Our goal is to link these datasets and reconstruct the original patients diagnostic pathways.

Pathways Linkage Scoring. In linking the patient’s longitudinal data, we consider a linkage score for each candidate pathway $Z = X \odot Y$, where higher scores imply higher chances for $X$ and $Y$ to match. Hence, $Z$ represents the same patient’s data evolution. In determining the linkage score, we consider two components. First, we evaluate the similarity in terms of shared subsequences (i.e. patterns) of the pathway $Z$ with the single patient pathways in the training set $T$. In fact, if we observe common temporal patterns between $Z$ and the sequences in $T$, this gives us a good indication that the pathways $X$ and $Y$ refer to the same user. We call this component the temporal agreement score for the candidate pathway $Z$. Second, patient’s attribute values may change over time. For example, a patient may receive a set of diagnoses that might change due to the disease evolution. Therefore, we consider a temporal disagreement score which measures this variation within the sequence. To evaluate this disagreement, we introduce a function that combines both the temporal and set-based between the observations within the pathway $Z$. This measure
is motivated by the concept of temporal locality, that is, if a time point a diagnose is observed, then it is likely that the same diagnose is observed in the near future. Then, our final linkage score $lScore(X,Y)$, combines both the diagnostic pathways temporal agreement and disagreement.

4  Experiment and results

Data. We conduct our experiments on two real-world datasets: MIMIC-III, and UCSD. MIMIC-III is a de-identified dataset which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. The data spans June 2001 - October 2012. UCSD dataset contains the patients admissions between January 2007 and December 2012 for a total of 41,730 patients in the hospital at UCSD. In our evaluations, we construct the diagnostic pathways of each patient using the information regarding the diagnoses and time recorded at hospital admission in each dataset. Since, we are interested in linking patient diagnostic pathways, we only keep the pathways that comprise at least two observations. As a result, we obtain 6,588 pathways and 18,850 pathways for MIMIC-III and UCSD respectively.

Evaluation Criteria. We perform 10-folds cross validation, where 9/10 of the data is used as a training set and 1/10 as a test case. For each patient in the test set, we randomly split his/her pathway across two databases $D_1$ and $D_2$. Then, for each patient partial pathway in $D_1$, our approach scores the pathways in $D_2$ using the linkage score and returns the top-$k$ pathways as a match. Different from the classic definition of True Positive and False Negative, we measure the accuracy of our solution as the ability of returning the real matching pathway in the top-$k$ results.

Table 1: Accuracy Results

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC-III</td>
<td>50.5%</td>
<td>66.4%</td>
<td>76.2%</td>
<td>89.0%</td>
<td>96.8%</td>
<td>99.5%</td>
</tr>
<tr>
<td>UCSD</td>
<td>56.3%</td>
<td>68.0%</td>
<td>80.1%</td>
<td>90.8%</td>
<td>95.5%</td>
<td>99.7%</td>
</tr>
</tbody>
</table>

Results. The accuracy results are reported in Table 1. The accuracy in linking patients is as high as 99% for both MIMIC-III and UCSD. Even when the top-1 path is retrieved, our solution leads accuracy higher than 50% demonstrating the effectiveness of our linking approach.

5  Discussion and Future Work

One of the major benefit of this approach consists in using the diagnostic pathways to represent the patients temporal evolution and in computing their similarity. In our setting, the shared information is only related to the patients diagnoses. Therefore, the privacy implications are limited, thus, more institutions would be willing to share this non-PHI information to effectively link patients data. Furthermore, the non-PHI data available in real scenarios are not limited to diagnoses but could also include more heterogeneous information such as: medications and lab tests. Therefore, as future research direction, we aim to extend our linkage framework to include all the temporal non-PHI data to further improve the linkage accuracy.

References


1Available upon registration: https://mimic.physionet.org
The Effect of a Large-Scale Mobile Health Emergency Project: A Telehealth Enabled Informatics EMS Intervention

Tiffany Champagne-Langabeer, PhD1; Michael Gonzalez, MD2; Diaa Alqusairi, PhD2; Adria Jackson, PhD2; Jennifer Mikhail, MPH1; James R. Langabeer II, PhD1
1The University of Texas School of Biomedical Informatics, Houston, Texas; 2Houston Fire Department, Emergency Medical Services

Introduction

Mobile health (mHealth) technologies have evolved in many areas of healthcare, but have not been widely adopted by emergency medical service (EMS) agencies. With almost half of all emergency visits initiated with ambulance transport, it is essential that mobile technologies better enable clinical and transport decisions. Technologies—including telehealth, health information exchange, and clinic scheduling—offer potential solutions in an environment where paramedics and physicians are increasingly resource-constrained with patients presenting with non-urgent complaints. Telehealth, specifically, shows potential given its ability to augment clinician productivity over distance through audio and video communication models. Within EMS, there is significant opportunity for transformation, given the historically high levels of emergency department transports for low-level acuity problems, resulting in unnecessary ED visits and overcrowding. We hypothesize an alternative model of technology-based emergency prehospital care will be effective at reducing costs and unnecessary hospital emergency department visits. This study describes one large regional mHealth program at the Houston Fire Department (HFD), Emergency Medical Services.

Objective

The objective of this research was to develop a mobile health technological solution for a large, urban EMS system, and to document the comparative effectiveness relative to traditional EMS care for primary care related incidents. The study was guided by two research questions:

1. How can telehealth and other mHealth technologies be implemented in a prehospital EMS environment?
2. What effects (in terms of clinical and economic outcomes) will such innovations produce for the Houston healthcare delivery system?

Methods

A project team was assembled of academics, informaticists, physicians, vendors, and medics to conceive a plan for a mobile health technology demonstration project in Houston, Texas. Technology vendors and partners were assembled, including major wireless carriers, tablet hardware provider, patient care record (PCR) systems provider, the regional health information exchange, and affiliated safety net clinics. A solution was designed to support transportation and treatment of non-emergent patients by more appropriate levels of care (such as a clinic instead of an emergency department). A schema of all interoperable web and mobile technology solutions was developed, and included telehealth, access to the regional health information exchange, clinic scheduling, and online taxi transportation solutions.
To study the effect of this program, we used an observational case-control study design. The intervention group received mHealth-enabled care and the control group was treated and transported per standard protocol. Primary outcomes were utilization (the % of ambulance transport to local ED) and productivity (“back in service time”, measured as cycle time differential from EMS start to finish). We compared the effectiveness analysis of the outcomes in each groups, comparing differences using Mann Whitney U median tests for statistical differences.

Results

The project initiated in December 2014 and has treated more than 6,000 patients (n=6,103) through February 2016. We found a significant reduction in medically unnecessary transports to the ED, with a 76% relative decrease in ambulance transports for the mHealth intervention group (74% ambulance transport for control group vs 18% for the mHealth program). This affected the median time from the initial EMS notification to its ability to back in service in order to respond to another 911 call. The median total response time was approximately 39 minutes for the mobile health patients, while median response for the control group was 83 minutes. This 44-minute median reduction in total response time per incident was statistically significant (Mann Whitney $P$<.001). This equates to approximately 2.1x greater utilization per day for the EMS unit when compared with the standard EMS control group. The units were able to return to service significantly faster as a result of less frequent trips to the ED, ensuring they can respond to other more emergent incidents. Table 1 summarizes the outcome differences.

Table 1: Outcome Differences

<table>
<thead>
<tr>
<th>Outcome Category</th>
<th>Measure</th>
<th>Control Group</th>
<th>mHealth (Intervention)</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilization</td>
<td>% Ambulance Transport to ED</td>
<td>74%</td>
<td>18%</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Cost of Care (Productivity)</td>
<td>Total Back in Service Time</td>
<td>83 (67) minutes</td>
<td>39 (34)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Discussion

A large-scale mobile health emergency project—incorporating telehealth and other innovative technologies—is effective at reducing unnecessary ED visits and putting EMS units back in service 2.1x faster. Telehealth, coupled with other technologies such as electronic scheduling and health information exchange, can enable transformation of the traditional EMS-based healthcare model. These results suggest broader implementation in other regions.

Conclusion

Mobile health technology enables an effective, alternative model of providing prehospital emergency care.

References

Mining and Visualizing Sequential Patterns in the Electronic Health Record: A Case Study for Asthma With and Without Mental Disorders

Elizabeth S. Chen, PhD1, Genevieve B. Melton, MD, PhD3, Mark Howison, MS2, Erik Knoll, BA2, Ashley Lee, MS2, Indra Neil Sarkar, PhD, MLIS1

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

Introduction

With the widespread adoption of electronic health record (EHR) systems, increased efforts are needed for adapting data mining techniques to transform EHR data into clinically actionable knowledge1. Association rule mining and sequential pattern mining are two such approaches that can be used to identify patterns in unordered and ordered sets of events, respectively2,3. In prior work, an open-source pipeline, incorporating the \texttt{arules} and \texttt{arulesViz} R packages, was developed to generate and visualize family history and substance use association rules from EHR data for pediatric asthma and adult epilepsy patients4,5. The goal of this study was to build upon this pipeline to identify sequential patterns using the \texttt{arulesSequences} R package, which has been previously used for predicting next prescribed medications6 and discovering comorbid neurodevelopmental disorders7. As a demonstration, a case study was performed for examining the development of comorbidities, including mental disorders, in patients with asthma.

Methods

A four-step process for data selection, data preprocessing and transformation, data mining, and interpretation was used8. First, ICD-9-CM diagnoses for adult asthma patients with at least two admissions were extracted from the Medical Information Mart for Intensive Care III (MIMIC-III; v1.3) database, which includes EHR data from Beth Israel Deaconess Medical Center from 2001 to 20129,10. Two datasets were generated that included diagnoses ordered by admission date for asthma patients with and without mental disorders. The 2015 Clinical Classifications Software (CCS) for ICD-9-CM that provides categorization schemes for diagnoses was leveraged to identify asthma and mental disorder diagnoses as well as to group all diagnoses into single- and multi-level categories11. The \texttt{arulesSequences} (v0.2-11) R package was used to generate sequences for both datasets at different levels of abstraction: (1) ICD-9-CM codes (e.g., 493.90 for “Asthma NOS”), (2) multi-level CCS categories (e.g., “Other asthma without status asthmaticus or exacerbation”), (3) single-level CCS categories (e.g., “Asthma”), and (4) ICD-9-CM chapters (e.g., “Diseases of Respiratory System”). To facilitate interpretation for this feasibility study, different filters were applied and visualizations of sequences were created (e.g., interactive Sankey and sunburst diagrams using the \texttt{googleVis} [v0.5.10] and \texttt{sunburstR} [v0.2.1] R packages). A preliminary comparison of sequences was performed to identify similarities and differences between asthma patients with and without mental disorders.

Results

The pipeline was implemented in R (v3.2.3) and PostgreSQL (v9.5.0). A total of 821 adult asthma patients, with 3.1±2.2 admissions and 14.5±7.1 unique diagnoses per admission, were identified in MIMIC-III. Sequences were generated at the four levels of abstraction for 539 (65.7%) asthma patients with mental disorders and 282 (34.3%) without. Figure 1 shows the proportion of common and unique sequences using single-level CCS categories for both datasets; an F-measure of 0.44 suggests differences in the progression of comorbidities in these subpopulations. Figure 2 depicts a visualization of a subset of unique sequences generated for asthma patients with mental disorders.

Discussion

The association between asthma and mental disorders (e.g., anxiety and mood disorders such as depression) has been long recognized12, but longitudinal studies are needed to ascertain temporal relationships of these conditions and contributions of social, behavioral, and familial risk factors over time13. This study is a first step towards an open-source framework for supporting such studies that may be generalizable to other conditions. Next steps include conducting formal evaluations of the generated sequences, incorporating algorithms to rank sequences and consider time intervals, examining the effect of different coding systems (e.g., ICD-9/10-CM and SNOMED-CT) and levels of abstraction (e.g., CCS for ICD-9/10-CM) on the number and quality of sequences, and applying the pipeline to EHR data (e.g., diagnoses, problem list, and social history) from different institutions and comparing sequences.
Figure 1. Number and Examples of Sequences (length a ≥ 2 and support ≥ 0.1) using Single-Level CCS Categories ([n%] = % of common and unique sequences; {A}{B} supp = Diagnosis A from an admission and Diagnosis B from a later admission with support value (i.e., proportion of patients) for respective datasets).

Figure 2. Sankey Diagram for Subset of 40 Unique Sequences (length a ≥ 2 and support ≥ 0.1) for Asthma with Mental Disorders (depicts flow of diagnoses from earlier to later admissions; link width = support; (n) = order in sequence)

References

Acknowledgments: This work was supported in part by National Library of Medicine grant R01LM011364.
Usability of an Automated Recommender System for Clinical Order Entry
Jonathan H Chen,1 Mary K Goldstein,2,3 Steven M Asch,1,4 Russ B Altman1,5
1 Department of Medicine, Stanford University, Stanford, CA, USA; 2 Geriatrics Research Education and Clinical Center, Veteran Affairs Palo Alto Health Care System, Palo Alto, CA, USA; 3 Primary Care and Outcomes Research (PCOR), Stanford University; 4 Center for Innovation to Implementation (Ci2i), Veteran Affairs Palo Alto Health Care System; 5 Departments of Bioengineering and Genetics, Stanford University

INTRODUCTION

The meaningful use era of electronic health records (EHR) presents learning health system opportunities to answer “grand challenges” in automatically learning clinical decision support (CDS) content. The current standard for distributing executable clinical expertise are human-authored order sets to reinforce best-practices. We previously developed a clinical order recommender system by automatically data-mining hospital EHR data. This approach aligns with established standards of care and is predictive of real physician behavior and patient outcomes, but its usability in a clinical decision workflow is unknown.

METHODS

As described previously, we extracted deidentified structured data for all inpatient hospitalizations from the 2009-2013 STRIDE clinical data warehouse. The data covers >74K patients with >11M instances of >27K items (medication, laboratory, imaging, and nursing orders, lab results and diagnosis codes). We built an order recommender based on this data, modeled on Amazon’s product algorithm using item co-occurrence statistics.

We built a simulated computerized physician order entry (CPOE) interface with open technologies including PostgreSQL, Python, Apache HTTP, and HTML/JavaScript. Our unique addition is an automated recommender (Figure 1), analogous to a “Customers Who Bought This Item Also Bought This…” function.

We developed clinical cases for common inpatient medical problems and initial trigger symptoms (with ICD9 code) of diabetic ketoacidosis (250.1), chest pain (786.5), neutropenic fever (288.0), and gastrointestinal bleed (578). Each case includes clinical notes to represent the patient’s history and physical exam. Lab values are only visible and change state if respective orders are entered (e.g., low hemoglobin reported only if a blood count is ordered). With each order entered, the recommendations update based on the accumulating patient information.

We recruited practicing physicians to use the interface to simulate admitting hospital patients. For the first case (diabetic ketoacidosis), the researcher guided participants through basic functions (data review and conventional order entry). For subsequent cases, the recommender function was activated. Participants completed a System Usability Survey (SUS) focusing on the dynamic recommender list. We organized qualitative observations and participant comments by thematic analysis.

RESULTS

Six board-certified Internal Medicine physicians participated in the simulated clinical order entry user-testing. All were local practicing physicians with 3-5 years of clinical experience, and thus familiar with the existing EHR order entry process. Their System Usability Survey (SUS) score median (IQR): 81 (76,90) translates to the 90.5th percentile (76.5,99.8) for usability relative to SUS normative evaluations. Table 1 includes major themes of comments and observations on the system. Table 2 includes design implications learned from initial issues.

DISCUSSION

We produced a clinical order entry interface and simulated cases that model changing patient states in response to user actions. This initial evaluation focuses on core usability, the commonly overlooked failure point for applied decision support, where as little as five participants identify the bulk of system usability problems. Physicians rated the integrated order recommender in the “excellent, grade A” category of usability and, more importantly, their qualitative comments and observations help to inform future system design.
### Category | Comment or Observation
---|---
Utility vs. Order Sets | • All participants used the pre-authored order set for the first example case (diabetic ketoacidosis)
|  | • No participant used case specific order sets once the recommender list was activated
Influence on Decisions | • “Recommender helps, but I’m deciding on my own. Reassuring though, that I am not alone.”
|  | • “I wouldn’t have ordered a CT scan, but I feel pressure to do so when it’s suggested.”
Trust of Suggestions | • “I know what to do, but I wish this safety net existed in real-life as I race between admissions.”
|  | • “I became comfortable with the recommendations until I realized a key antibiotic was missing.”
Interpretability of Scores | • “I like the prioritization of orders, not buried amongst useless ones.”
|  | • “The actual numbers don’t inform my decisions unless they are extreme.”

Table 1 – Participant comments and researcher observations on the recommender system.

All participants used the pre-authored diabetic ketoacidosis order set which is *required* in real-world settings to order an insulin drip. For subsequent cases, no participant used (and most were not even aware of) pre-authored scenario specific order sets from the hospital EHR. In contrast, every participant used multiple system recommended orders for all cases, suggesting a more seamless integration with their thought processes.

We observed no gross medical management errors of commission, though there were arguably errors of omission. Only one participant ordered prophylactic antibiotics (i.e., ceftriaxone) for a patient with a variceal GI bleed, which notably was *not* on the computer generated recommendation list. At the same time, most participants did not order filgrastim for neutropenia, despite it prominently featuring in the recommendation list. Multiple participants argued that they focus on acute stabilization and defer second order decisions until they can speak with a specialist. In fact, participants complained about seeing recommendations for procedures that only a specialty consultant would perform / order. These reinforce that a recommender system (and even conventional order sets) remain decision *aids* that a human decision maker still must consider in the context of their individual patient.

<table>
<thead>
<tr>
<th>Initial Design</th>
<th>Issue / User Response</th>
<th>Design Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long scrollable list</td>
<td>Distracted by poring over the entire list (100)</td>
<td>Avoid a long list of suggestions</td>
</tr>
<tr>
<td>Multiple stats reported</td>
<td>Ignored as too much to digest and interpret (e.g., prevalence, PPV, RR, P-value)</td>
<td>Simplify numerical annotations (e.g., high, medium, low)</td>
</tr>
<tr>
<td>Items sorted by single dimension (e.g., PPV)</td>
<td>Hard to scan list for a specific item when sorted by arbitrary number</td>
<td>Organize suggestions into traditional categories (e.g., labs, meds, etc.)</td>
</tr>
<tr>
<td>Separate lists sorted by PPV or RR</td>
<td>Confused why items on two lists. PPV preferred even though more mundane.</td>
<td>Consider interleaving suggestions from different selection methods</td>
</tr>
<tr>
<td>All items eligible for all users</td>
<td>Objected to specialty order recommendations (e.g., cardiac catheterization, endoscopy)</td>
<td>Consider limiting recommendations based on user role</td>
</tr>
</tbody>
</table>

Table 2 – Apparent issues in initial recommender design with respective implications learned from simulation.

Overall, practicing clinicians were receptive to an automated recommender integrated into their order entry decision process. With confidence in the core usability of such a design concept not disrupting clinical workflow, further evaluation of such data-driven clinical decision support systems will determine whether they can actually improve the quality and efficiency of medical decision making in a closed-loop learning healthcare system.

**REFERENCES**

Using Hospitalization and Suicide Mortality Data to Identify Subpopulation of High Suicide Risk via Survival Modeling
Kun Chen, PhD¹, Fei Wang, PhD², Robert Aseltine, PhD³
¹Department of Statistics, University of Connecticut. Storrs. CT.
²Department of Healthcare Policy and Research. Weill Cornell Medical College. NYC. NY.
³University of Connecticut Health Center. Farmington. CT.

Abstract

Objectives: To demonstrate a statistical survival analysis for identifying subpopulation at high risk for suicide using statewide hospitalization and suicide mortality data. Methods: Eight years of claims data (2005-2012) on hospitalizations due to intentional self-injury from the Office of the Connecticut Medical Examiner and the Connecticut Hospital Inpatient Discharge Database were combined with mortality data by suicide, to study the survival time of patients after their most recent hospital admissions due to unsuccessful suicide attempt. For each subject, his/her historical records were aggregated to the time of his/her most recent hospitalization that did not result in death. We consider a “follow-up” study of these subjects, starting from their most recent admission times that did not result in death until death due to suicide or until the end of the study at 12/31/2012. From suicide prevention perspective, this is exactly the cohort that is at high risk of suicide-related death and is exactly at the time that the prevention efforts can be readily applied. A cause-specific Cox’s proportional hazard model, coupled with regularized variable/interaction selection techniques, was used to build a parsimonious and interpretable model, to study the associations between the survival times and the potential predictors constructed from the hospitalization records. The out-of-sample predicted survival probabilities from the Cox model were used to identify high-risk groups from the study cohort. Results: Several important predictors and their interactions were identified to be associated with the survival time after most recent unsuccessful suicide attempt. The predicted individual survival probability is demonstrated to be a valid risk measure of suicidal death. High-risk group with several folds of increased risk was identified based on the out-of-sample predicted survival probabilities, and several insights were gained on the unique characteristics of the high-risk group. Policy Implications: Data capturing hospitalization for suicide attempts and suicide deaths can inform prevention activities by identifying high-risk subpopulation to which resources should be allocated.

Methods

Eight years of claims data (2005-2012) on hospitalizations due to intentional self-injury were combined with mortality data on death by suicide over the same period, to study the survival time of patients after their most recent hospital admissions due to unsuccessful suicide attempt.

Using historical hospitalization data from the Office of the Connecticut Medical Examiner and the Connecticut Hospital Inpatient Discharge Database, we identified all the suicide-related hospital admissions during an eight-year period from 2005 to 2012. A fabricated subject identifier was created to identify subjects with multiple admissions, by matching gender, age, ethnicity, birth date, and residential zip code (We remark that the exact identifier indicating multiple visits of the same subject was not available). For each subject, his/her historical admission records were aggregated to the time of his/her most recent admission time that did not result in death. The aggregated admission records include demographical variables (such as gender, age, ethnicity, etc), summary statistics of visit histories (such as number of suicide-related admissions, average length of stay in hospital, etc), and numerous indicator variables created from diagnostics codes, procedure codes and discharge status codes (ICD9). 164 subjects who died during their only hospital admission were removed from the study, and we also limited the study to subjects under 70 years old at their most recent admission times due to unsuccessful suicide attempt. This results in totally 15,246 unique subjects during the 2005-2012 periods. As such, we consider a "follow-up" study of these subjects who were admitted to hospitals due to suicide attempt, starting from their most recent admission times that did not result in death. From suicide prevention perspective, this is exactly the cohort that is at high risk of suicide-related death and is exactly at the time that the prevention efforts can be readily applied. We then used the CT Suicide Mortality Database to find out whether these subjects have died due to suicide until 12/31/2012. It can be viewed that the 15,246 subjects at risk were “followed” until death due to suicide or until the end of the follow-up study at 12/31/2012. (As the complete death records in CT were not available, we were unable to identify deaths due to other causes; nevertheless, the potential bias is controlled, as we have limited the study to subjects under 70 years old.). We conducted a statistical survival analysis to model the associations between the survival times and the potential predictors constructed from the admission records. Specifically, a cause-specific Cox’s proportional hazard model, coupled
with regularized variable/interaction selection techniques, was used to build a parsimonious and interpretable predictive model. The final model is selected based on cross validation. High-risk subgroup can then be identified based on estimated individual survival probabilities from the fitted model. To objectively access the predictive power of our approach in identifying high-risk subgroup, we repeatedly split the data to 80% training and 20% testing randomly; the survival model was fitted using the training data and then used to identify the high-risk group, defined as 5% of the patients in the testing data having the lowest predicted individual 5-year survival probabilities. The estimated 5-year survival probability of the high-risk group and that of all the patients in the test data were then calculated. This procedure was repeated 500 times. Results from this analysis indicate whether the identified highest risk cohort indeed have much higher risk of suicidal death than that of all attempters.

Results
Predictors included patient demographics, primary and secondary diagnosis codes (both ICD 9 and MS-DRG), and primary and secondary procedure codes. We considered all the main-effect and two-way interaction terms as candidate variables. Altogether 18 separate risk factors and combination of risks were identified. Our out-of-sample random splitting procedure showed that while the 5-year mortality rate among the entire previous attempters was 3.5%, the highest risk cohort identified by our predictive model, comprising 5% of all attempters, had a 5 year mortality rate of 14.3%. The identified 12 separate risk factors and combination of risks can indeed differentiate the high risk group from all patients hospitalized for a suicide attempt (Table below). White men and older patients were significantly more likely to be in the high risk group; those whose prior attempt involved hanging or suffocation were 5.6 times more likely to be in the high risk group. Several conditions and comorbidities related to the prior attempt were also predictive of being at high risk for suicide, including acute alcoholic intoxication (ICD 9 303), alcohol intoxication in the context of poisoning by sedatives and hypnotics (ICD 9 967), alcohol intoxication in the context of open wound of other and unspecified sites except limbs (ICD 9 879).

<table>
<thead>
<tr>
<th>All the reported ratios are significantly different from 1 at significance level &lt;0.0001, based on an exact test of the ratio of Poisson rates.</th>
<th>All Patients</th>
<th>High-risk group</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.42</td>
<td>0.95</td>
<td>2.2</td>
</tr>
<tr>
<td>White</td>
<td>0.72</td>
<td>0.99</td>
<td>1.4</td>
</tr>
<tr>
<td>Age (Mean)</td>
<td>35.3</td>
<td>48.0</td>
<td>1.4</td>
</tr>
<tr>
<td>Acute alcoholic intoxication</td>
<td>0.15</td>
<td>0.38</td>
<td>2.6</td>
</tr>
<tr>
<td>Black</td>
<td>0.09</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>Medicare (vs. Private Ins)</td>
<td>0.09</td>
<td>0.27</td>
<td>2.9</td>
</tr>
<tr>
<td>A fluid filled sac within the cerebrum</td>
<td>0.02</td>
<td>0.12</td>
<td>6.8</td>
</tr>
<tr>
<td>Poisoning and toxic effects of drugs</td>
<td>0.08</td>
<td>0.31</td>
<td>3.9</td>
</tr>
<tr>
<td>Alcohol/Drug abuse or dependence w/o rehabilitation therapy</td>
<td>0.01</td>
<td>0.04</td>
<td>3.5</td>
</tr>
<tr>
<td>Other persons seeking consultation</td>
<td>0.01</td>
<td>0.07</td>
<td>9.2</td>
</tr>
<tr>
<td>Hanging/Suffocation</td>
<td>0.02</td>
<td>0.13</td>
<td>5.6</td>
</tr>
<tr>
<td>Personal history of malignant neoplasm</td>
<td>0.01</td>
<td>0.06</td>
<td>5.5</td>
</tr>
<tr>
<td>Poisoning by other central nervous system depressants and anesthetics</td>
<td>0.02</td>
<td>0.09</td>
<td>4.9</td>
</tr>
<tr>
<td>Symptoms, signs, and ill-defined conditions</td>
<td>0.17</td>
<td>0.30</td>
<td>1.8</td>
</tr>
<tr>
<td>Other psychosocial circumstances</td>
<td>0.12</td>
<td>0.27</td>
<td>2.2</td>
</tr>
<tr>
<td>Poisoning by sedatives and hypnotics</td>
<td>0.07</td>
<td>0.17</td>
<td>2.3</td>
</tr>
<tr>
<td>Acute alcoholic intoxication &amp; Open wound of other and unspecified sites except limbs (ICD 9 879)</td>
<td>0.00</td>
<td>0.03</td>
<td>15.1</td>
</tr>
<tr>
<td>Acute alcoholic intoxication &amp; Poisoning by sedatives and hypnotics</td>
<td>0.01</td>
<td>0.10</td>
<td>10.6</td>
</tr>
</tbody>
</table>
A Framework for Visual Tracking of Risk and its Drivers in Monitoring Patients Susceptible for Cardiorespiratory Instability

Lujie Chen, MS\textsuperscript{1,2}, Gilles Clermont, MD\textsuperscript{3}, Marilyn Hravnak, PhD\textsuperscript{4}, Michael R. Pinsky, MD\textsuperscript{3}, Artur Dubrawski, PhD\textsuperscript{2}

\textsuperscript{1}Heinz College and \textsuperscript{2}School of Computer Science, Carnegie Mellon University, and School of Medicine\textsuperscript{3} and School of Nursing\textsuperscript{4}, University of Pittsburgh, Pittsburgh, PA

Abstract

Monitoring cardiorespiratory instability (CRI) in hospitalized patients is challenging, primarily due to the heterogeneity of patterns of CRI risk escalation and the diversity of the involved risk drivers. We present a framework for producing interpretable visualizations of CRI risk estimates obtained from multiclass machine learning models. Our approach can be used as a tool for “aerial view” of patients’ risk characteristics as well as support online monitoring and real time tracking of CRI risk.

Introduction

Cardiorespiratory instability (CRI) denotes a physiologic decompensation event observed among hospitalized patients. Its early recognition and intervention can improve outcomes\cite{1}. However, detection and forecasting of CRI is a challenging task, primarily due to the heterogeneity of CRI risk escalation patterns\cite{2} and to the diversity of risk drivers\cite{3}. Existing bedside monitoring systems are capable of recording large amounts of fine-grained physiological data. In current practice, however, these data are rarely used to track and characterize patients’ risk trends in real-time. We propose a novel visualization framework to support tracking of CRI risk escalation patterns and risk drivers over time. It relies on assimilating large amounts of multi-parameter vital signs (VS) data using a multi-class classification model, and visualizes the resulting risk estimates for easy bedside interpretation.

Method

In our previous work\cite{2}, we described a method to characterize and phenotype patient’s CRI risk escalation patterns using Group Based Modeling. We further extended this work by developing a multi-class Random Forest classification model to track patient’s group status over time using the numeric features extracted from VS from recent past. The output of the model is a multi-dimensional vector whose components estimate the probability of the patients to follow a particular CRI risk escalation pattern, given their VS signals observed up to the current time. Here we also include an independent multi-class classification model to track the CRI risk drivers in six dimensions (high heart rate [HR], low HR, high respiratory rate [RR], low RR, low blood oxygen level [SpO\textsubscript{2}] and abnormal blood pressure [BP]), and the resulting multi-dimensional estimation vector can be interpreted similarly.

For each patient, we project the series of multi-dimensional estimated probability vectors onto 2-dimensional polygons, as shown in Figure 1. The vertices of each polygon represent unique categories in our classification models, while each projected data point corresponds to the tracked patient state at a certain point of time. The mapping between a multi-dimensional vector and the coordinates in the plot follows the formula:

\[
x^t = \sum_{i=1}^{k} x_{0i} P^t_i \quad y^t = \sum_{i=1}^{k} y_{0i} P^t_i
\]

Here, $P^t_i$ represents the probability that a patient’s state belongs to the $i^{th}$ category at time $t$ according to the model. Each pair $(x_{0i}, y_{0i})$ represents the coordinates of the vertex of the projection polygon corresponding to the $i^{th}$ category, whereas $(x^t, y^t)$ are the coordinates of the projected data point. We only map points with risk levels higher than a user-determined threshold based on the output of a binary classifier deciding whether a patient is at risk.

This approach permits us to map the model-based estimation of the patient’s current state in a multi-dimensional space into a point inside a 2-dimensional convex polygon for intuitive interpretation. For example, if the model estimates that a patient’s current state belongs to a particular category with high CRI likelihood, the corresponding data point will be rendered in proximity of the vertex representing that category. If, however, the estimation is split
between two distinct categories, then the data point will be located along the line connecting the corresponding category-specific vertices. It also follows that if the output of the model shows a high level of uncertainty about the category assignment, i.e. the estimated probability distribution is close to uniform, then the projection will be close to the center of the polygon. By varying temporal resolution of data to be visualized, we can modulate the presentation to visualize the dynamics of the patients’ CRI risk evolution over time between longer and shorter horizon perspectives.

**Results**

Figure 1 illustrates example visualizations obtained by tracking a specific patient’s CRI risk with respect to their risk escalation pattern and drivers. Figure 1(A) plots all tracked states during this patient’s entire length of stay whenever their estimated risk level exceeds a threshold of 0.5. The plot on the left summarizes the patient’s risk escalation patterns among three categories (“persistent high”, “early onset” and “late onset”), while the plot on the right traces the estimated risk drivers among the six categories. Contour lines depict density distribution of data. We can see that this patient’s most common risk escalation pattern was the “late onset” class with an occasional drift towards the “persistently high”. The primary risk drivers for this patient observed during their stay were low SpO₂ and high RR. Figure 1(B) illustrates the view when this tool is used for online monitoring in which only the last two hours of the patients estimated risk states are shown. Blue arrows illustrate the temporal sequences of the state transitions. Supplemental animation videos are available from this link: [http://lujiec.github.io/visual_tracking/](http://lujiec.github.io/visual_tracking/)

![Figure 1. Example visualization of (A) retrospective view of one patient’s CRI risk trajectory estimation and CRI risk drivers throughout their stay, and (B) trace of these estimates during the most recent 2 hours of their stay.](image)

**Conclusion**

We present a framework that can be used to obtain meaningful and interpretable summarization and visualization of the evolution of monitored hospitalized patients’ patterns of CRI risk escalation and the corresponding CRI risk drivers. Our framework can be used as a tool for either retrospective analysis conducted in order to profile and interpret the changes of a patients’ risk characteristics over longer intervals of time, or for real-time tracking to provide the clinical team with a fine-grained and communicative risk estimation information in order to devise CRI monitoring, mitigation, and remediation strategies tailored specifically to each patient.

**References**

Bulk Learning on EHR Data

Po-Hsiang Chiu, PhD\textsuperscript{a}, George Hripcsak, MD, MS\textsuperscript{a}
\textsuperscript{aDepartment of Biomedical Informatics, Columbia University, New York, NY}

Introduction

A central computational task in data-driven disease phenotyping is to use the variations of clinical concepts inferred from the electronic health records (EHR) to stratify phenotypic cohorts, each of which represents a clinical condition or a potential disease subtype that goes beyond historical disease definitions. Patients are essentially represented by clinical traits such as medication history and laboratory measurements. Phenotyping using statistical learning methods\textsuperscript{1} may seem to reach a higher level of automation when compared to the conventional rule-based approach; however, predictive analytics is not without its own challenges in the creation of training data since both the tasks of feature engineering and gold standard labeling require significant human intervention. To reduce the overhead of data annotation and achieve a higher level of automation, we adopt a hierarchical learning approach based on the ensemble learning paradigm\textsuperscript{2}, using infectious diseases as the domain of study, to construct abstract features that represent the shared clinical traits among the target infections. Statistical models such as logistic regression can then be built from within the abstract feature space of low dimensionality, thereby reducing the demand for labeled data.

Method

Bulk learning, as the name suggests, is a simultaneous training procedure for multiple clinical conditions. The key idea is to use diagnostic codes (e.g. ICD-9s) readily available in the codified EHR as surrogate labels, out of which we can train a hierarchical model with levels of feature abstraction that captures the common denominator of the target conditions, collectively referred to as the bulk learning set. Clinical evidence suggests that different infectious diseases share degrees of similarity in antibiotic prescriptions, various laboratory tests, among other clinical factors, all of which can potentially serve as high-level phenotypic measures for separating cohorts with different infections. Through the learned abstract features representing importance weightings of phenotypic measures of interest, the labeling of disease cases for the entire bulk learning set can be realized by training statistical models from within the abstract feature space. Since abstract feature space has a significantly lower dimension compared to the raw feature space spanned by patient attributes extracted from the EHR data, we postulate that a relatively small annotated dataset is only required for stratifying disease cohorts, which would otherwise have a substantially higher demand of labeled data considering the size of the raw feature set that can be used to model these phenotypic measures. An example three-tier hierarchical learning architecture based on stacked generalization\textsuperscript{3} is depicted in Figure 1: i) Base or level 0 consists of a set of classifiers, each of which is modeled to express a high-level phenotypic measure relevant to the target infections. Each measure is represented by the feature set extracted from Medical Entity Dictionary (MED) developed at NewYork-Presbyterian Hospital\textsuperscript{4}, a semantic network for medical concepts. ii) The level-1 tier is a meta-classifier that takes probabilistic scores from base classifiers and a set of indicators combined as the abstract feature set, where each indicator represents the absence or presence of data for the corresponding phenotypic measure (e.g. not all patients had intravenous chemistry tests during their clinical visits); up until this level, ICD-9 codes are used as surrogate labels. iii) The level-2 tier inherits the probabilistic output from level l coupled with the ICD-9 codes together as the abstract feature set while using annotated labels for model training.

Results and Discussion

In the experiment, we used the clinical data repository from Columbia University Medical Center and examined 100 clinically different infectious conditions. For simplicity, $\ell_2$-regularized logistic classifiers are used at both the base and the meta levels. In predicting the ICD-9 as surrogate labels, each base classifier has different predictive strength due to their varying degrees of clinical relevance. The phenotypic measure involving microbiology tests in general has the best predictive strength. The level-1 classifier consistently outperforms individual base classifiers, which can be explained by the diversity in predictions at the base level, a crucial condition for ensemble learning to work well. In the model evaluation with the gold standard, we annotated 83 clinical cases. Under this small annotation set, the ICD-9-modulated level-2 model reaches an AUC of 0.78. However, the model tends to result in false positives due to the inconsistent labeling between the ICD-9 and the gold standard; overcoming this performance shortage is an ongoing effort. Additionally, we experimented on alternative feature abstractions such as the ICD-9-modulated level-1 model, which combines the level-1 features with the ICD9; the performance tradeoff with its level-2 counterpart depends on the size of labeled data. Figure 2 illustrates the relative performance (in AUC) of the global
The level-1 model in reconstructing the ICD-9 labeling (through predictions), where the model was trained on the aggregated level-1 data across the bulk learning set (hence global to all diseases), followed by predicting individual diseases as the evaluation; the horizontal axis denotes the codified diseases. The empirical result indicates that the global level-1 model approximates the ICD-9 labeling well (with a grand mean AUC at 0.897), which is promising for using such hierarchical learning method for large-scale, high-throughput phenotyping in the near future.

**Figure 1.** Three-tier learning architecture with two levels of feature abstractions over 4 phenotypic models.

**Figure 2.** Sorted performance in AUC for the global level-1 model across the bulk learning set.

**Acknowledgment**

This work was funded in part by National Library of Medicine grant LM006910.

**References**

Patient Perspectives on Designing an Engaging Patient-Powered Research Network Patient Portal: Crohn’s and Colitis Foundation of America (CCFA) Partners
Arlene E. Chung, MD, MHA, MMCi,1-2 Maihan B. Vu, Dr.PH, MPH,1 Jessica Burris,3
Kelly Myers,3 Michael D. Kappelman, MD, MPH1-2
University of North Carolina at Chapel Hill,1 Carolina Health Informatics Program,2
Vanderbilt School of Medicine,3 Atomo, Inc.3

Background. Patient-powered research networks (PPRN) have the potential to improve health behaviors and outcomes, yet initiating and sustaining engagement is challenging. The overarching goal for CCFA’s PPRN is to promote both research participation and disease management. In order to build a PPRN patient portal that prioritizes the needs of its members who have inflammatory bowel diseases (IBD), we sought to better understand patients’ needs and preferences for what they feel to be essential components and features to facilitate and sustain engagement with the PPRN patient portal with particular focus on patient-generated health data (PGHD). As PPRNs are an emerging concept, little is known about patient’s preferences and information needs for PPRN patient portals and what features are considered the most important for self-management and engagement. This study identifies design principles for creating an engaging and impactful PPRN patient portal. We postulate that some of these principles could be useful for developing PPRN portals across other clinical conditions.

Methods. We conducted a two-phase qualitative study to elicit patient perspectives on what would make the PPRN patient portal engaging and useful for self-management and monitoring. In Phase 1, seven 90-minute focus groups involving 62 patients with IBD from two states were conducted with recruitment purposefully focused on including minority IBD patients. The focus group results informed the Phase 2 cognitive interviews, which included thirteen 60-minute phone interviews with patients from 11 states. Topics discussed in the group and individual sessions included barriers and facilitators to engagement with the PPRN and focused on the various aspects of PGHD and potential resources and tools to facilitate self-management within the PPRN’s patient portal. Individual interview participants were also shown mockups of PGHD displays including patient-reported outcome data. All focus groups and interviews were digitally recorded, transcribed verbatim, and analyzed in ATLAS.ti 7.5. From the combined data, thematic categories were derived and codes were grouped into emergent themes and relationships. We used these themes to determine design principles for the PPRN portal.

Results. Participants from the seven focus group discussions (n=62) were mostly middle-aged, female, and self-identified as White. 61% had Crohn’s disease and reported an average 21 years since diagnosis. Almost all focus group participants had a cell phone (mostly smartphones), Internet access on their cell phones, and had computers or tablets with Internet access. Individual interview participants (n=13) had an average age of 46 years with 93% being female and the majority identified as White. Approximately half of participants had Crohn’s disease and reported an average 16 years since diagnosis. All individual interview participants had a cell phone and computers or tablets with Internet access.

The four predominant themes that emerged through inductive coding techniques were: (1) The Impact of Knowing; (2) Participation Barriers and Challenges; (3) Engagement and Collaboration; and (4) Customizable Patient Portal Features and Functionalities. The design principles informed by each of the major themes are detailed in Table 1.

Discussion. Partnering with patients to inform the design principles for the development of CCFA Partners PPRN patient portal was critical for creating an engaging and impactful portal. Patient insights and perspectives on what features and functionalities are important for self-management and engaging in research provided invaluable information to inform the design of the PPRN portal.

Table 1: Design Principles for the CCFA Partners PPRN Patient Portal

<table>
<thead>
<tr>
<th>Themes</th>
<th>Issue</th>
<th>Design Principle</th>
</tr>
</thead>
</table>
| The Impact of Knowing           | Patients desire information about various treatment options & to learn the experience of others to help inform their own health and treatment decisions. Patients were motivated to participate because the knowledge gained in research studies would benefit both society and the individual patient (finding a cure, prevention for others, etc.) | Create a section on the portal to:  
• Provide resources and information about treatment options based on the current evidence base  
• Disseminate research findings from studies conducted with PPRN data in a format that patients can easily understand  
• Disseminate information on what studies are being currently conducted |
| Participation Barriers and Challenges | Patients are unsure about the credibility of IBD information online (both type and amount). Patients desire to be able to judge whether or not information is from a valid source. | Show data about what studies are being conducted using patient data and how their individual data is being used for each study conducted within the PPRN. Provide information and sources for any information posted on the portal, including conflicts of interest or industry relationships, so patients can determine the validity of the information. Provide a way for patients to choose who can use their PGHD & clearly state that patient data will never be sold. |
| Data security and protection was a concern. | PPRN portal is secure & privacy/security information should be visible and available on the PPRN portal website. |
| Engagement and Collaboration | Patients desire to know that someone from the PPRN is paying attention to them and interested in them as patients. | Provide daily check-in survey about overall health and IBD symptoms. Provide intermittent emails to update membership about new features and results from PPRN studies. |
| | Patients feel more likely to sustain engagement and participation if they feel invested in the PPRN’s products and outcomes. | Obtain continuous input from patient governance members and membership on what features and products are most impactful to them. Have an easy way for patients to submit feedback/suggestions. |
| | Personal invitations from their providers and some type of periodic check-ins for support and encouragement would facilitate engagement. | Create personalized invitations to join the PPRN from patient’s providers and provide a way for providers to support/encourage patients via messaging using the portal. |
| | Patients feel more likely to track with wearables/smartphone apps during flare-ups and when stressed, but also viewed continuous tracking as an opportunity to track how they were doing & how their body was doing. | Provide an easy mechanism to connect and disconnect various wearables/smartphone apps to track PGHD, which allows patients to pick which devices/apps are the most meaningful or helpful to them. |
| | Patients cited the need for a feedback loop for how their PGHD is being used for research. | Display how PGHD is being used and by whom, & provide patients with results from studies that use their PGHD. |
| Patient Portal Features and Functionalities with Focus on PGHD | Patients desire the ability to track trends (health status & IBD symptoms) over time. Patients specifically want the ability to pick what measures to compare and graph. Individual trends over time were more important than comparing to other patients. Participants desired both an overall & detailed view of their PGHD. | Provide a graphing feature that allows patients to select the measures & time frame to graph. Provide individual and comparative data displays for patient-reported outcomes data and wearable/app data. Provide an overall view of data trends over time but allow for click-to-expand on individual data points. |
| | Patients desire access to personal health information from electronic health records. | Using HL7, enable data exchange from patient portal/personal health record accounts. |
| | Patients desire to customize or change features to meet individual needs, especially for PGHD displays. | Provide patients an easy way to click and select features they wish to use within the MyMeasures section, which displays various types of PGHD and other health data. |
| | Patients felt that using a stoplight color scheme for PGHD displays was universally understood and helpful. | Display symptom and severity data in stoplight color scheme. |
| | Participants would like a dashboard of all their data and information in one place, and this dashboard should help communicate this information to their providers. | Create a dashboard using color-coding (stoplight colors) when possible to display wearables/app data, patient-reported outcomes, and disease activity. Allow for sharing of dashboards with providers. |

**Acknowledgements.** This research was supported by CCFA, NIH P30 DK034987, PCORI, and GlaxoSmithKline. Dr. Chung also receives support from NIH NCATS 1KL2TR001109.
A Qualitative Study of User-Desired Personal Health Record Functionality: Impact of Age on Desired PHR Functionality.

Martina A. Clarke¹, Marilyn Sitorius¹, Tom A. Windle¹, Ann L. Fruhling², Tamara L. Bernard¹, John R. Windle¹

¹University of Nebraska Medical Center, Omaha, NE, ²University of Nebraska – Omaha, Omaha, NE

Introduction

The Personal Health Record (PHR) is intended to support patients’ access to data, clinical summaries, preventive care, educational materials, and medication reconciliation. It is one of the core requirements for Meaningful Use Stage 2. The PHR aims to improve medication adherence, self-management of disease, and the patient provider communication. Despite the potential benefits of PHRs, adoption has been poor, in part due to usability issues. We believe that by understanding the needs of users, the use and utility of the PHR will increase. A core hypothesis of this project is that desired PHR functionality will, in part, be determined by the age of the patient.

Methods

To determine how patients want their PHR to function, a qualitative study was conducted using the grounded theory approach. Following Investigative Review Board approval; patients were recruited to participate in a focus group to determine their thoughts about the PHR and how it fit into their care. Inclusion criteria for the study comprised of patients currently scheduled for clinic follow ups at the Cardiovascular Center at the University of Nebraska Medical Center, 19 years of age or older, and willing to give consent. The study was designed to recruit participants from a range of ages and economic diversity. Economic diversity was achieved by using zip code data from the 2010 census.

Focus groups consisting of 1-5 participants began July 2015 and sessions lasted approximately one hour. A moderator gave a scripted explanation of the study, and both audio and video recordings were obtained. The moderator asked participants questions concerning their use, if any, of the PHR, how they would like to use the PHR, and features they would like the PHR to provide. Our interdisciplinary team (2 physicians, 1 nurse and 3 informatics professionals) asked the participants relevant follow up questions, such as, questions concerning securing health information, questions related to communication with healthcare providers, device preference for accessing the PHR, and usefulness of the after visit summary (AVS). The audio and video files and transcripts of the recorded interviews were reviewed, and analyzed by two independent coders. The codebook was created using a thematic approach. Themes were revised until coders came to consensus and any disagreements between the two coders were settled by a third independent reviewer. Data analysis was carried out using NVivo©. We determined the frequency, intensity and convergence of the coded information. This study was funded in part by AHRQ grant: Optimizing the Electronic Health Record for Cardiology Care (R-01 HS022110-01A1).

Results

Focus groups including 39 individuals (19 males and 20 females) were conducted. Twenty-six participants were white, six Hispanic, five black, and two American Indian. Twenty-nine participants lived in an urban location and ten participants lived in a rural location. Twelve out of 39 participants resided in zip codes with low mean income. Responses were classified according to age groups: 19-39 (6 participants) referred to as young adults, 40-64 (8 participants) referred to as middle-aged, and >65 referred to as seniors (25 participants). Five out of six young adults were users of a PHR, three out of eight middle-aged participants were non-users of a PHR, and fourteen out of twenty-five seniors were users of a PHR.

Several common desired features of the PHR among the age groups were identified. These included the ability to not only access laboratory and test results, but displayed so they could see normal values and trends. In particular, they desired the physician’s interpretation. All age groups requested links to credible sources about their diseases. Most stated a desire to view their physician’s notes, or impressions, in lay terms. Participants within all age groups communicated electronically with their physician, either through their PHR or via e-mail. Most felt that it would be valuable to communicate with their provider before scheduled appointments in order to provide a list of questions and to provide their current medication list. Participants found the use of passwords cumbersome to remember and keep track of and were open to alternative authentication methods for accessing their PHR.

Several differences in the desired use of the PHR among age groups were identified. Young adults had interest in having access through the PHR to social groups with patients who shared similar diseases. When asked if
the after visit summary (AVS) was useful, young adults did not find the AVS very useful because the information found in the AVS was also available in the PHR and preferred to have access to an electronic version of the AVS. Young adults preferred to access their PHR through an app on their smart phones. Young adults also requested to be able to update information on the PHR, and to have the ability to communicate with providers through text, email, and through secure messaging within the PHR.

Reasons middle-aged participants (40-64) mentioned for not using a PHR includes: trust and security when using the Internet, being computer illiterate, not knowing if messages were received by their provider, and perceived frustrations with the functions of the PHR. Middle-aged participants did not mention access to social media as a priority. Preference for contacting provider was through phone call or email but no text messaging. Middle-aged participants, who were willing to use the PHR, preferred a PC or tablet to smartphones and larger font display. The paper-based AVS was more likely read by middle aged participants and stored for record keeping purposes.

Seniors (>65) also did not mention the need or use for social engagement with other patients. Most seniors found parts of the AVS useful but preferred paper copies. Seniors preferred communication with their providers through phone call, email, and in person. A unique feature requested by seniors is to give caregivers access to health information online. Seniors, who used the PHR, preferred a PC or tablet and also stated the need for improved accessibility in order to use the PHR, by the use of a larger font.

Discussion
Several critical desired features were identified in this study that could improve adoption of PHRs. Young adults high use of the PHR for convenience and preference for a smart phone over other modalities. A survey by Pew Research Center found that 85% of young adults (18-29) and 79% percent of adults (30-49) owned a smartphone. This could be why young adults in our study preferred accessing their PHR from their smartphone. Middle aged and seniors’ use of tablets could be because tablet technology is less intimidating, presenting a more simplified interface when accessing the Internet. A study by Jayroe and Wolfram found that older adults had an overall positive experience when using a tablet to complete search tasks. Seniors mentioned interest in having caregivers gaining access to their health information. The reason for this requested feature by seniors may be because of their dependence on caregivers for their health management.

This study is novel in its categorization of desired functionality of PHR by participants’ age groups. These results suggest that providing patients with an optimized PHR is not a ‘one size fits all’ endeavor. The results demonstrate that patients of different age have different PHR functionality requirements and preferences on how they would like information presented to them. These results could inform providers on the information patients seek and could improve the communication providers have with their patients during clinic visits. These results could also inform vendors of patient’s desired PHR functionality for the development of an improved PHR.

Conclusion
This study should inform the design and implementation the PHR to improve usefulness of PHRs and hence increase the use and satisfaction of the PHR. While this study included a diverse group of patient’s further study with a larger sample size is necessary to confirm these results. Future research should assess a PHR based on the results. Additionally, future research should include how caregivers are included in the communication between patient and provider and how caregivers may gain access to patients’ health information.

References
Variation in EHR Documentation across Primary Care Providers

Genna R. Cohen, BS ¹, Charles P. Friedman, PhD¹,²,³, Andrew Ryan, PhD¹, Julia Adler-Milstein, PhD¹,³

¹School of Public Health, ²Department of Learning Health Sciences, ³School of Information
University of Michigan, Ann Arbor, MI

Introduction: Substantial public and private funding has been invested to increase the adoption of electronic health records (EHRs) (1) because of an expectation that EHR use will lead to higher-quality, lower-cost care (2). EHRs do not, however, directly alter states of disease or health: their value depends on how they are used, and EHRs typically offer providers substantial flexibility in how to use them. This means that providers often have multiple choices for how to document the same type of information. When information is consistently documented in the EHR it is easier for clinicians to find and act on relevant details of the patient’s history. Consistently documented information also allows for tools that are designed to automatically extract information from the patient chart, such as clinical decision support functionality that alerts users to best practice management of medical conditions as well as reports to manage patient populations. However, variation in documentation likely undercuts these benefits.

Variation in EHR documentation is likely explained by a mix of “appropriate” and “inappropriate” factors. For example, variation within physicians across encounters is most likely to be appropriate since it should be primarily driven by patient characteristics. Variation is also likely to be mostly appropriate across offices within geographic regions, and across regions. The remaining variation, variation across providers within offices, is therefore the most likely to be inappropriate. Because providers of the same specialty in the same office see a similar mix of patients (3), this variation would reflect different preferences across providers and a lack of consensus about how best to document comparable information within an EHR (4). It is therefore critical to single out variation at this level. Data from a national, commercial EHR vendor are utilized to answer the following research questions:

(1) How much variation in EHR documentation occurs across PCPs for 15 categories of clinical documentation, and which categories reflect the greatest variation?

(2) For categories with the greatest variation, what proportion of variation in EHR documentation remains after accounting for variation across states, provider organizations (POs), and offices?

Results help inform ongoing efforts to optimize EHR use, and leverage these tools to support improvements in the quality, safety, and efficiency of healthcare delivery.

Methods: De-identified EHR log data from a web-based EHR vendor provide a record of granular clickstream data for all encounters during the month of June 2012 in ambulatory primary care offices. Data were grouped into 15 overarching clinical documentation categories, such as collecting vitals or updating or reviewing the problem list. Every element of EHR documentation is tied to a patient visit and a user ID. The data also capture nested relationships, linking users to their office, a parent PO, and state. The final analytic sample included 170,332 encounters led by 809 PCPs nested in 237 offices in 76 POs in 27 U.S. states.

For each encounter, 15 binary indicators capture activity within each clinical documentation category by anyone in the office. For each PCP, this yields a proportion of encounters with completed documentation (e.g., any activity) for each of the fifteen clinical documentation categories (i.e. fifteen outcome measures per PCP). Demographic data create a categorical variable for PCP primary care specialty type as well as PCP, office, PO and state identifiers.

To answer the first research question about variation across PCPs, the standard deviation of the proportion of encounters for which a given clinical documentation category is completed categorizes tasks according to their variation level (based on the standard deviation). Clinical documentation categories with a standard deviation above 25% were considered High Variation.

To answer the second research question a multi-level linear regression model for each of the “High Variation” clinical documentation categories calculates the proportion of variation that remains after accounting for variation explained by state, PO, and office. The outcome variable is the proportion of encounters for which a given clinical documentation category is completed for PCP i nested in office j nested in PO k nested in state l. The model includes a single fixed effect parameter that captures PCP specialty because there are likely differences in the distribution of patients treated by each type of PCP and therefore their EHR documentation. State- and PO-specific effects (random intercepts) are also included to capture variation in documentation occurring between units at each of these levels. This explained variation is expressed as a ratio to total variation for each level (state, PO, and office). The main variable of interest is the remaining variation across PCPs that is not explained by specialty, office, PO, or
state, also expressed as a ratio to total variation. Statistical significance in the ratio of unexplained variation is a signal of inappropriate variation because it reflects variation across providers in the same office. Bootstrapped standard errors were used because the theoretical distribution of the test statistic for this ratio is unknown.

Results: After analyzing the proportion of PCPs’ encounters in which each clinical documentation category was completed and the standard deviation across PCPs, five clinical documentation categories were considered High Variation: Updating the Patient’s Social History; Conducting a Review of Systems; Reviewing and Discussing Documents; Updating the Patient’s Problem List; and Assessing and Diagnosing the Patient. These tasks were completed in 32.5% - 66.3% of providers’ encounters on average, with high standard deviations of 28.5%-36.9%.

Multilevel models for the five High Variation clinical documentation categories reveal statistically significant variation across states for two clinical documentation categories: Reviewing and Discussing Documents (7.1% of variation explained, p<.05) and Updating the Patient’s Social History (4.0% of variation explained, p<.05). There was significant variation across POs for all but one clinical documentation category (8.3%-17.8% of variation explained, p<.01); the exception was Reviewing and Discussing Documents during the Visit (5.0% of variation explained, p=.21). There was statistically significant variation across practices for all but one clinical documentation category (13.5%-19.0% of variation explained, p<.01); the exception was Reviewing and Discussing Documents during the Visit (9.9% of variation explained, p=.08). Finally, the majority of variation for every clinical documentation category was the remaining, unexplained variation, not accounted for by variation across states, POs, or practices. Documentation of Social History had the least unexplained variation (62.2%, p<.001). The clinical documentation category with the most unexplained variation was documentation of Reviewing and Discussing Documents during the patient’s visit (78.1%, p<.001).

Discussion: Optimal EHR documentation requires pursuing a balance between the needed flexibility to accommodate differences across patients and a recognition that too much flexibility allows for inappropriate variation that compromises downstream uses of EHR data (5). The majority of variation in documentation in a national sample of PCPs exists across users in the same office, which is likely a signal of inappropriate variation. Indeed, the specific High Variation clinical documentation categories are those for which there are multiple places in the EHR to complete documentation. This creates opportunity for PCP preferences to result in variation. For example, a Review of Systems is often structured as a component of an admission note covering the patient’s organ systems. If the elicited information leads to the identification of a problem or diagnosis, that information could reasonably documented in the Problem List, as part of the Assessment & Diagnosis, or in all three categories. The observed, unexplained variation in completing these three clinical documentation categories may therefore be the result of different PCP preferences about the relative benefits of documenting in different sections of the EHR (4).

While the data do not allow for a definitive explanation of why certain clinical documentation categories are completed more variably than others, it provides direction for future research in this area. Specifically, additional analysis should explore variables at the provider- and office-level that may explain additional variation, such as the number and type of users participating in a given encounter by role (physician, nurse, etc). These types of variables would not necessarily reflect differences in patient needs, and would provide additional insight into the provider and staff work patterns that may contribute to variation. In the interim, EHR vendors and physician practices may want to consider developing clear guidelines and norms regarding High Variation clinical documentation categories to ensure that information which is entered in the EHR may be easily retrieved and used for high-value purposes.

References
Increased Understanding of Safety Concerns Experienced by Hospitalized Patients and their Families through Real-time Reporting in MySafeCare App

Sarah A Collins RN, PhD\textsuperscript{1,2,3}, Brittany Couture\textsuperscript{1}, Elizabeth Lilley M.D., M.P.H.\textsuperscript{1,4}, Ann DeBord-Smith M.D., M.P.H.\textsuperscript{1,4}, Frank Chang\textsuperscript{2}, Esteban Gershanik MD, MPH, MMSc\textsuperscript{1,3}

\textsuperscript{1}Brigham and Women’s Hospital, Boston, MA; \textsuperscript{2}Partners Healthcare Systems, Boston, MA; \textsuperscript{3}Harvard Medical School, Boston, MA; \textsuperscript{4}Center for Surgery and Public Health, Brigham and Women’s Hospital, Boston, MA

Introduction
Care concerns experienced by hospitalized patients and their families are seldom reported in the absence of a safety event causing patient harm. A lack of reporting may stem from various influences, such as patients or families feeling uncomfortable in disclosing negative feedback to the clinical community or a lack of clear communication paths and coordination of follow-up. These influences combine to greatly limit our understanding of patient and family perspectives on safety threats and decrease our ability to intervene in real-time at a patient level. Moreover, analyzing patterns of safety concern data permits system-level recognition of targets for practice and process change. Engaged, activated patients experience fewer adverse events, have shorter lengths of stay, incur lower healthcare costs, and have a lower likelihood of pursuing malpractice litigation in the event of patient harm.[2–4] Data-driven mitigation of potential safety events and real-time analysis of trends based on patient data could transform hospitals into accountable, patient-centered learning health systems.[5] To better understand, mitigate, and track safety concerns from the patient and family perspective we developed and implemented MySafeCare, a web-based application that allows hospitalized patients and families to electronically submit and categorize safety concerns in real time with the option to remain anonymous. The user-centered design and evaluation process for MySafeCare has elucidated a number of important sociotechnical findings related to safety concerns from the patient and family perspective that have impacted our iterative design and implementation approaches. This podium presentation will describe: 1) the iterative user-centered development and implementation of MySafeCare, 2) quantitative analysis of submissions captured through MySafeCare, and 3) qualitative analysis of the unique perspectives of patients and families related to reporting safety threats identified and explored in this project.

Methods
We performed rapid iterative development and piloting of the MySafeCare tool, including the patient facing application (see Figure 1) and the administrator and clinician dashboard. System development included 7 steps: 1) stakeholder engagement, 2) iterative user-centered design, 3) extraction and prioritization of requirements, 4) wireframes and prototyping, 5) development, 6) testing, 7) version revisions based on testing and end-user feedback. The tool is deployed at the unit level to facilitate unit-based views of the clinical dashboard. Results reported here are for the initial 7 months of the pilot study. Data from all 15 months of pilot study (through fall 2016) will be included in the presentation. Version 1 was piloted for 3 weeks on a Vascular Surgery unit at Brigham and Women’s Hospital (BWH) during March and April 2015. Version 2 was implemented for a 6 month trial from May - October 2015 on the Medical Intensive Care Unit (MICU) and the Oncology Units (ONC) at BWH. Version 3 went live on MICU and ONC in February 2016 and will expand to medicine and surgical units later in 2016. We will present the counts and types of concerns reported from all versions and study units. Additionally, we will compare data submitted to MySafeCare to other safety data at our institution before and during the study period on our study units. These comparisons illustrate the novel information that MySafeCare captures. To better understand the unique perspectives of patients/families related to reporting safety threats we conducted interviews with patient and families and collected surveys on patient willingness to engage in their care and safety. These data are triangulated with the free text narrative of submitted safety concerns to define a conceptual model of the current state of patient and family safety reporting and to identify how data captured in MySafeCare can contribute to a learning health system.

Results
Based on our initial requirements gathering, the first prototype of MySafeCare was web-based so that patient/family users could “Bring Your Own Device” and did not require a log-in, allowing for patient/family users to remain anonymous (if desired) and to report a concern at anytime during their hospital stay. These features persisted as important to end-users throughout our testing and evaluation. Additional features include options for categorizing safety concerns, submitting a narrative of the concern, ranking the severity of the concern, remaining anonymous or self-identifying, and submitting concerns to a clinical dashboard that is viewable by the Nurse and Medical Director,
Patient Relations, or the patient’s care team. While disclose of identity could occur through details provided in the narrative, it is at the discretion of the patient/family to enter that information and that has not occurred to date.

Six concerns and 5 compliments have been submitted to MySafeCare: 5 from patients and 6 from family members. Among the concerns, 5 (83%) were submitted anonymously. Concern categories included care plan, communication, and infection. Per unit, this equated to 1 concern submitted every: 7 days on the Vascular Unit, 65 days on Oncology Unit, and 131 days on the MICU. In comparison, on average the Patient and Family Relations Department receives 1 complaint every: 33 days on Vascular Unit, 29 days on Oncology Unit, and 34 days on MICU. While counts are relatively low they were comparable to patient complaints received by Patient and Family Relations at our hospital during the same time period the prior year, and the content of concerns were useful and unique. Interestingly, a patient perceived harm related to intravenous infection control was in reality a best practice for changing intravenous line caps that the patient had not been educated about. The majority of patients/families interviewed were unaware of the administrative ‘chain of command’ of the hospital and paths for reporting a safety concern.

**Discussion**

MySafeCare application represents the next-generation of actionable measurement tools that provide a voice to, and data of, patient’s safety concerns. We know that patients have different comfort levels disclosing negative feedback to the clinical community – often referred to as “white coat syndrome”.[1] Hospitalized patients’ level of comfort in disclosing negative feedback poses a great challenge to the integrity and completeness of patient-reported safety data. MySafeCare successfully captured anonymous concerns and identified unique infection control concerns from the patient and family perspective. The notion that some patient and family concerns were due to a lack of knowledge of best practices for infection control illustrates the need to capture, track, and attend to the patient perspective to not only mitigate harms but mitigate the perception of harm when none has occurred. Patients and families lack of knowledge of the “chain of command” available for reporting safety concerns and the value of reporting concerns before they result in safety threats or harms illustrates a major gap in patient engagement strategies and the need for sociotechnical approaches to bridge this gap. We observed great variation in use of MySafeCare across clinical units. We are studying the trends and typical users on each unit (ie., family are typical users in ICU; Oncology patients often openly communicate concerns with Oncologist) to better understand if care unit, disease state, and prior experience with the health care system could impact comfort in safety reporting and the need for an anonymous, electronic reporting tool. We will report on lessons learned related to the effective implementation approaches and how these vary based on the patient population per unit. Finally, we will summarize our triangulated data of patients/families’ unique perspectives related to hospital safety.

**Acknowledgements:** This work was funded by AHRQ 1P30HS0235335 Making Acute Care More Patient Centered.

**References**


Helping Patients with Breast Cancer Decide about Clinical Trials Participation

Prudence Dalrymple, PhD, MS¹, Lisl Zach, PhD,² Amy Leader, DrPH, MPH³, Laura Austin, MD³, Sarah Hegarty, MPhil⁴, Anna Quinn, MPH ³, Anett Petrich, RN³, Massimo Cristofanilli, MD,⁵ Russell Schilder, MD³, Ronald Myers, PhD³

¹College of Computing & Informatics, Drexel University, Philadelphia, PA; ²Informatics Insights, LLC, Philadelphia, PA; Department of Medical Oncology, Thomas Jefferson University, Philadelphia, PA; Division of Biostatistics, Thomas Jefferson University, Philadelphia, PA; ³Lurie Cancer Center, Northwestern University, Chicago, IL

Introduction
Treatment advances depend on clinical trials (CT), but numerous barriers to trial participation exist, including limited patient engagement in shared decision-making (SDM).¹ Previous studies have shown that decision aids (DAs) have a positive effect on the deliberative process of health decision-making.² We developed an online Decision Counseling Program (DCP)³ as an interactive DA to facilitate patient education and preference clarification. The qualitative research presented here is part of a larger study aimed at using the DCP to promote SDM between patients with breast cancer and their oncologists about CT participation.⁴ The goal of this sub-study was to elicit from patients their experience with the DCP. Findings provide insights and illustrate the challenges of implementing an interactive DA in clinical practice.

Method
Seven women who were newly diagnosed with breast cancer presented at an office visit for consultation with an oncologist about treatment. Immediately preceding the consultation, each woman met with a trained decision counselor, who obtained informed consent, administered a baseline knowledge survey, and used the DCP to review printed information on CTs and to complete a preference clarification exercise on CT participation. The online DCP generated a one-page summary form that displayed the patient’s preference for or against participating in a clinical trial; this summary form was given to the patient and placed in her medical chart for reference in an ensuing patient-oncologist clinical encounter. Six of the seven clinical encounters were successfully audio-recorded; one provider declined to be recorded. Approximately 30 days after the visit, structured interviews were conducted with the patients by phone and audio-recorded. These interviews ranged from 10 to 25 minutes in length (mean = 17:24 minutes; median = 15:22 minutes). Transcripts of the clinical encounters and the patient interviews were analyzed to identify key themes and patterns using open coding, which was pre-tested for intercoder reliability. This technique ensures that explanations developed from the data are verified repeatedly to protect against researcher bias.

Results
Seven women were enrolled in the sub-study. Participants ranged in age from 47 to 85 years; five women were white and two were nonwhite; two were married or living with a partner; and all had at least a high school education. Based on the results of the DCP, the derived preference of five out of seven women was to join a clinical trial; one patient had a neutral preference and the other preferred not to join a trial. One physician engaged in a short dialogue about the DCP
process itself, and another mentioned the study but didn’t refer to the materials; one nurse practitioner explained that education about clinical trials is “a good thing” even if no trial is available at the time. Although each participant was given a DCP summary form, analysis of the clinical encounter transcripts showed that the form was not mentioned in the encounters.

Analysis of the post-visit interview transcripts indicated that most participants did not recall the specific components of the DCP session, but did remember the educational materials and found them useful. They particularly noted that the materials improved their knowledge about CTs, especially that CTs normally do not include a “no-treatment” placebo control group. The women indicated that at the time of the DCP session, they felt “overwhelmed” because it was so soon after their diagnosis. This situation may explain why their recall of the session was incomplete. One woman who did recall the preference clarification exercise said “[I found it] helpful because I am sort of a visual person.”

Discussion
Our analysis of the transcripts from the clinical encounters and post-visit interviews are consistent with the findings of previous studies that patient knowledge increased from baseline to 30 days. However, it appears that completing the DCP session did not directly impact SDM about CT participation, perhaps because none of the participants was eligible for a clinical trial at the time of the sub-study encounter. However, the DCP may still be a valuable tool to prepare women who are eligible for a CT to engage in SDM, since it helps patients clarify their preferences and supports knowledge acquisition. It may also be the case that providing guidance to patients and training for physicians on how to use information generated by the DCP will be necessary to facilitate SDM about CTs.

Conclusion
Our results contribute to greater understanding of the complex challenges of integrating SDM into the clinical encounter. While several studies of the DCP have been published, this is the first time, to our knowledge, that the clinical encounter following the DCP session has been audio-recorded, thus affording insights into its impact on SDM reported in the patients’ own words.

References
Assessing the potential risk in drug prescriptions during pregnancy

Ferdinand Dhombres MD PhD, Vojtech Huser MD PhD, Laritza Rodriguez MD PhD, Olivier Bodenreider MD PhD
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Introduction

Over eighty percent of pregnant women in the United States are prescribed at least one drug during pregnancy. The U.S. Food and Drug Administration (FDA) regulates the labeling of drug products and has established five risk categories for drug use during pregnancy. This classification was introduced in 1979. New FDA regulations (June 30, 2015) deeply revised the pregnancy and lactation labeling by eliminating these categories and replacing them with narrative summaries describing the risk of the drug and supporting evidence [1]. In a recent study on Medicaid data, 40% of pregnant women were dispensed at least one medication from categories D or X, for which there is positive evidence of human fetal risk [2]. In this work, our objective is to assess the potential risk in drug prescriptions during pregnancy, with respect to the new FDA standard. A secondary objective is to contrast risk assessment between the old FDA categories and the newly introduced narratives.

Materials and methods

**Acquiring reference risk categories.** As a proxy for the new FDA standard, we used the “pregnancy recommendations” from a reference textbook (Briggs, 10th ed. 2015) [3]. For each ingredient, it provides the level of risk (contraindicated, high risk, moderate risk, low risk, probably compatible and compatible with pregnancy), the source of evidence, if any (human or animal data), and other information as appropriate (trimester, dose and drug association restrictions). For the original FDA categories, we used an older version of the same textbook (Briggs, 8th ed. 2008), where each ingredient was associated with one of the 5 categories used at that time (A, B, C, D and X). When an ingredient was associated with more than one category (e.g., to account for risk variation based on dose, length of exposure, or associated comorbidities), we used the highest risk category. For vitamins, however, we took the lowest risk category, because prescriptions were generally within the Recommended Dietary Allowance (RDA).

**Acquiring and processing prescription data.** We analyzed patient-level, de-identified claims data of a privately insured population of 159.7M patients from 2003 to 2014 provided by the IMEDS Research Lab. We relied on procedure codes for delivery to identify pregnant women (13 CPT (Current Procedural Terminology v4) codes covering all vaginal deliveries and caesarean sections). We considered a period of 270 days prior to delivery or C-section for drugs dispensed during pregnancy. We used the RxNorm API to relate drugs from claims data to the reference. We derived the risk and supporting evidence associated with each drug, taking the highest risk in case of multi-ingredient drugs. We restricted our analysis to systemic drugs, because topical drugs generally pose a much lower risk. We counted prescriptions by category, using the new standard (level of risk and source of evidence) and the old FDA categories. We also contrasted the two standards. Two OB/GYNs (FD and LR) reviewed the top 50 of each category to ensure the reliability of the results.

Results

A total of 3,741,743 pregnant women were selected, to which 19,654,083 prescriptions were dispensed (15,815,624 for systemic drugs). The level of risk was defined using the classification extracted from Briggs (for the old and new risk categories) for 14,719,736 prescriptions (93%).

**New risk categories and supporting evidence.** Overall, 40.2% of the prescriptions were “compatible” with pregnancy and 1.2% were “probably compatible”. The prescriptions were contraindicated in 2.8%. There was a potential risk in 8,191,485 prescriptions (55.6%). For prescriptions for which the risk was quantified, the risk was low (37.6%), moderate (1.5%), and high (0.03%). For 60.8% of prescriptions, however, the risk was not quantified. Overall, evidence based on human data is available for 91.85% of all prescriptions. For “compatible” and “contraindicated” prescriptions (i.e., 43.0% of all prescriptions), the evidence was always based on human data, as defined in the Briggs recommendations. For prescriptions with a potential risk, the source of evidence was “human data” in 87.8%, “limited human data” in 10.7%, and “only animal data” in 1.49%. Only for a small fraction of the prescriptions with a potential risk (0.005%) was the evidence based on limited data, irrelevant animal data, or no data at all.
Comparison with the old FDA risk categories. (The definitions of the categories are adapted from Briggs 8th ed.)

- Almost all prescriptions originally categorized as A (i.e., controlled studies in women fail to demonstrate a risk to the fetus) are now listed as compatible with pregnancy.
- Similarly, all prescriptions originally categorized as X (i.e., with positive evidence of fetal risk that clearly outweighs any possible benefit) are now listed as contraindicated. Differences are generally due to trimester, dose and drug association restrictions.
- Prescriptions originally categorized as D (i.e., with positive evidence of fetal risk but benefits from use during pregnancy may be acceptable despite the risk) are now associated with a potential risk in 92.3% (low risk for 36.3%) and are contraindicated in 2.3%.
- Prescriptions originally categorized as C (i.e., either animal studies indicate a fetal risk, and there are no controlled studies in women, or no studies are available) are now listed as compatible with pregnancy in 46.6%, are associated with a potential risk in 51.2% (low risk for 29.1%) and are contraindicated in 0.5%.
- Finally, prescriptions originally categorized as B (i.e., either animal studies do not indicate fetal risk and there are no controlled studies in women, or animal studies have shown an adverse effect, but controlled studies in women failed to demonstrate a risk) are now listed as compatible or probably compatible with pregnancy in 41.2%, and are associated with a potential risk in 58.7% (low risk for 18.1%). None of these are contraindicated.

Discussion

Findings. This investigation demonstrates the feasibility of assessing the potential risk in drug prescriptions during pregnancy from a large claims dataset using RxNorm and the Briggs reference, with finer-grained recommendations compared to the old FDA categories, as well as stronger evidence. It had already been demonstrated that pregnant women are commonly prescribed drugs associated with fetal risk [2]. However, supporting evidence was not reported. Our results show that the proportion of prescriptions without reliable human data evidence was small (8.15%). Interestingly, in the Briggs reference, human data evidence is available for only a third of the ingredients associated with a potential risk. In contrast, in our cohort, there is human data evidence for 87.8% of the prescriptions for drugs with potential risk.

Limitations and future work. This preliminary investigation did not take into account recommendations for specific trimesters of pregnancy, which we will address in future work. This is of particular importance since risk may significantly differ over time. For example, *misoprostol*, an abortive drug, is contraindicated during pregnancy, but it is also widely used near term for labor induction [4]. Dose can impact the level of risk as well. For example, vitamin A is compatible with pregnancy under the Recommended Dietary Allowance (RDA), but contraindicated above the RDA. Several common drugs have a dose-related risk (e.g., *aspirin*, *fluconazole* and most vitamins), but the drug products commonly prescribed during pregnancy generally correspond to lower doses (e.g., *baby aspirin*, *multi-vitamin supplements*). However, for complex risk assessment (comorbidities, co-prescriptions, precise dose, duration of exposure), claims data may be insufficient.

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine, and by the French Gynecology and Obstetrics Association (Collège National des Gynécologues et Obstétriciens Français) and the Philippe Foundation and by the Reagan-Udall Foundation for the FDA (IMEDS Research Lab).

References

Integrated Workflow Capture in an EHR Conversion: Standardizing on Best Practice Methods

Brad Doebbeling, MD, MSc1,2; Matthew Burton, MD3; David Kaufman, PhD1; Karl Poterack, MD4; Mac McCullough, PhD1; Adela Grando, PhD2,4; Rick Helmers, MD3; Tim Miksch, MBA5
1School for the Science of Healthcare Delivery & 2Department of Biomedical Informatics, Arizona State University, Phoenix, AZ; 3Mayo Clinic, Rochester, MN; 4Mayo Clinic Arizona, Phoenix, AZ, USA

Introduction:
Widespread implementation and meaningful use of health information technologies (HIT) has been hailed as a crucial step towards a learning health care system with safer, higher quality care. Implementation of the meaningful use policies has fostered major investment in electronic health records (EHRs). However, this rapid adoption has drawn attention to many challenges. EHRs are often insufficiently designed, tested, or integrated with clinical workflow in mind.1-4 In addition, the system transformation efforts involving EHR implementation nationally have often been costly with little or no documented improvement in interoperability, unless a conceptually-driven comprehensive transformation is planned.3,5 They have also been associated with various unintended-consequences- workarounds, adverse events, inefficiencies, and patient, provider and staff frustration.3-5

The productive use of HIT is determined by the extent to which it informs clinical decision-making and provides cognitive support for optimal clinical workflow. Little consensus exists around in-depth approaches for capturing clinical workflow prior to implementation of an EHR. Most organizations either do little to document the current state, or rely on private consulting groups or the EHR vendor to make these assessments. However, failure to do this important groundwork sets up the organization for potential failures in implementation, poor provider, staff and patient satisfaction, and inability to use these data for subsequent optimization and redesign efforts. HIT implementation or redesign that fails to consider workflow changes can lead to increased medical errors.3,7

Experts claim that the fit of HIT into clinical workflow is the most important basic necessity for safe, effective, and efficient HIT use; workflow can be assessed and described at many levels and dimensions that are often in conflict.5,7,9

Many capture, analysis, and representation approaches have been employed for evaluating complex clinical workflows.8,9 These fall into clusters of research that address, at best, a few dimensions of clinical work using theories, methods, and frameworks specific to a single discipline.10 Thus it is challenging to systematically describe work that involves highly complex, distributed, collaborative, continuously evolving, interruptive, cognitive/informational and physical elements.1,2

Methods:
The Mayo Clinic has invested significant resources to develop a rich electronic ecosystems for clinical information management. Much effort by IT professionals and providers has gone into eliciting and understanding the needs of clinicians and manifesting those needs as functionality and capabilities in numerous niche applications designed, implemented, and continuously evolved to facilitate workflow to enable high quality care. The Mayo Clinic is embarking on a large scale EHR implementation to replace multiple first generation EHRs. This undertaking has been named the Plummer Project in honor of Henry Plummer, MD, who drove Mayo’s transition in 1907 from clinic and physician records kept in ledger books to a patient-centered medical record. This project will involve migration from current, disparate EHR platforms (e.g., GE Centricity, Cerner) to a future state of a single, primary EHR (Epic Systems).

The convergence of a single, enterprise-wide EHR will enable if not necessitate the convergence of many operational processes across Mayo sites. This change will have a profound impact on the workflow of providers and staff, given the number and degree of niche, siloed applications in existence to meet the specific job duties of very specialized groups. Migration to a newly integrated EHR will be completed throughout the enterprise by 2018, giving an unusual and time-limited opportunity to document in-depth the current heterogeneous state before it is replaced. Documentation will not only serve an archival purpose but also provide invaluable “prior state” data that can be used to characterize existing pre-implementation processes in sufficient depth to reveal what has been lost in the conversion, as well as opportunities for further optimization, tool development and research and development. It will also serve as a basis to empirically assess the extent of workflow change, through process and outcome measures for both providers and patients. Additional benefits include informing data semantics for historical records and research initiatives that leverage longitudinal clinical data.

The ROOT project is both an acronym (Registry Of Operations and Tasks) and the name of Henry Plummer’s assistant Mabel Root, who played a major role in the transition to a patient-based medical record. This project begins in April 2016, to capture and archive existing HIT-mediated workflow with the current 3 EHRs (GE Centricity and 2 versions of Cerner) prior to conversion to another EHR (Epic) in all 4 of its health systems. It is a systematic effort to capture and curate a substantial portion of the current electronic ecosystem. Through the ROOT project, we will conduct a multi-method approach to reveal an in-depth understanding of work components (e.g., individuals, groups, contextual factors, information resources), as well as the interactive behavior and dependencies between the components that influence HIT-mediated performance. We will employ a comprehensive, multi-disciplinary, and multi-method Clinical Workflow Capture and Analysis methodological framework that draws on
numerous expert and evidence-based best practices and conceptualizes a more holistic and integrated view of clinical workflow consisting of interdependent, formal task structures and co-constructed, often informal task structures. The findings from various disparate, yet complementary approaches can be normalized through formal task-structures that explicitly define “who” does “what” (activities), “when”, “where”, and “how” (sub-activities) by employing “which resources or tools”, consuming, altering, or producing “what information”, and in “what relation” to other tasks or entities formally expressed in a Work Domain Ontology and visually depicted as Activity Network Diagrams. The informal task-structures are further elicited to express the actions of individual entities interacting dynamically in a complex activity system through mechanisms of coordination, awareness of co-workings, and qualitative descriptors that preserve the richness of findings from distinct methods.

A series of interrelated, interdisciplinary projects are planned for 1) capturing and archiving a massive multisite, multiprovider observational, interview and video dataset of EHR, clinical workflow, and contextual information; 2) discovering workflows from event logs generated by various information and communication systems; and 3) creating a repository of work data that serve as digital signatures, or patterns in an organization to understand common practices, deviations, and need for intervention and monitoring. A modified Delphi Panel of over 60 Mayo Clinic clinical experts involved in the Plummer project will further develop and rank a comprehensive list of clinical areas, providers and important workflows to be studied by the project. We will survey asynchronously clinical and operations leaders using the over 200 information systems that will be replaced and integrated by the Plummer project.

Results:
Mayo and ASU informaticists, clinicians and researchers have been partnering to define the methods and technologies needed to capture a sufficiently comprehensive snapshot of the current state. We will present results from the first six months, challenges and solutions to the barriers encountered.

Discussion:
The approach outlined here is important as a necessary step in preparation for the system transformation associated with an EHR implementation or conversion. We intend to use the ROOT project as a “reference implementation” of such tools, techniques, and methods to set the standard for future workflow analysis when implementing, optimizing, or enhancing HIT. The methodologies and tools applied to the ROOT project will be customized for use at Mayo and published as an established and authenticated standard for the organization. Without collecting and archiving such data prior to implementation, it will always be difficult to address questions of whether the patients, providers and processes are improved. Furthermore, such a strategy will help ensure the implementation is successful, that post-implementation optimization is well-informed, and that questions about functionality that may have been lost can be addressed. Starting with appropriate collection and curation of data, analyses may help to develop and disseminate software and analytic tools to streamline the process of modeling clinical workflow. We anticipate a program of R&D may hasten innovation, guide implementation of more efficient delivery systems, and help transform the health care system.

References:
Delivering High Quality Birth Certificate Data from an EHR

Jeffrey Duncan, PhD¹, Catherine Staes, BSN, MPH, PhD²
¹Utah Department of Health, Salt Lake City, UT; ²University of Utah, Salt Lake City, UT

Introduction

Birth certificates are one of the core public health data sets and are used for public health surveillance, policy development, and research.¹ Birth certificate data are typically reported to state public health departments by hospitals using the 2003 US Standard Birth Certificate.² In practice, hospital birth clerks, typically employees in the Health Information Management or Medical Records departments are tasked with abstracting maternal and child information onto standard worksheets, then keying that information into a web-based Electronic Birth Registration System (EBRS) to report to public health. As healthcare facilities increasingly adopt electronic health record (EHR) technology, there is growing interest in automating the capture and reporting of birth certificate information. The National Center for Health Statistics (NCHS) has led efforts to develop and test methods for reporting birth certificate medical information from EHRs, resulting in a standard Integrating the Healthcare Enterprise (IHE) profile known as Birth and Fetal Death Reporting enhanced (BFDRe).³ BFDRe has been tested by EHR and EBRS vendors during interoperability showcases but has not yet been implemented operationally.

As the BFDRe standard matures and jurisdictions prepare for implementation, it is important to understand the quality, limitations, and transformations required to automate birth certificate reporting from an EHR. Our primary objective was to assess the accuracy of information extracted in the BFDRe CCD compared to that reported on birth certificates for select maternal and child data elements in order to ascertain readiness for automated birth reporting. An additional objective was to understand the strengths and limitations of automated birth reporting and to develop strategies to ensure birth certificate data quality. Table 1 shows the birth certificate data elements that were compared, organized by source of information.

Table 1. Birth certificate data elements in this report organized by source

<table>
<thead>
<tr>
<th>Source</th>
<th>Birth Certificate Data Element and Item Number on 2003 Birth Certificate Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenatal Record</td>
<td>Date of first prenatal care visit (Item 29a)</td>
</tr>
<tr>
<td></td>
<td>Gestational diabetes (Item 41)</td>
</tr>
<tr>
<td></td>
<td>Total number of prenatal visits (Item 30)</td>
</tr>
<tr>
<td></td>
<td>Date of last normal menses (Item 39)</td>
</tr>
<tr>
<td>Labor and Delivery</td>
<td>Fetal presentation at birth (Item 46C)</td>
</tr>
<tr>
<td></td>
<td>Method of Delivery (Item 46D)</td>
</tr>
<tr>
<td></td>
<td>Antibiotics received by mother during labor (Characteristics of labor and delivery) (Item 45)</td>
</tr>
<tr>
<td></td>
<td>Induction of labor (Characteristics of labor and delivery) (Item 45)</td>
</tr>
<tr>
<td></td>
<td>Augmentation of labor (Characteristics of labor and delivery) (Item 45)</td>
</tr>
<tr>
<td>Newborn Record</td>
<td>Obstetric estimate of gestational age (Item 50)</td>
</tr>
<tr>
<td></td>
<td>Assisted ventilation following delivery (Item 54)</td>
</tr>
<tr>
<td></td>
<td>Assisted ventilation &gt; 6 hours (Item 54)</td>
</tr>
<tr>
<td></td>
<td>NICU Admission (Item 54)</td>
</tr>
</tbody>
</table>

Methods

Utah Department of Health (UDOH) worked with University of Utah Health Care (UUHC) to evaluate the quality of 12 selected birth certificate variables documented in its Epic EHR. In December 2015, UUHC implemented an interface with Epic’s Stork module to create HL7 CDA-standard Labor and Delivery Summary (LDS) documents for a sample of 60 recent births. The same information was collected at Sibley Memorial Hospital (SMH) in
Washington, D.C. and compared to birth information reported to the District of Columbia Department of Health (DCDOH). Information from the LDS was compared with: a) information previously reported on birth certificates, and b) a reference standard created by two birth certificate experts from UDOH and two from DCDOH who independently reviewed each of the 60 records in their respective jurisdictions and adjudicated differences.

**Results**
The availability and accuracy of information automatically extracted from the EHR varied widely depending on source of data and location. Table 2 shows preliminary results of comparisons between the Extract group and the EBRS group compared to the Reference Standard for both Utah and DC.

**Table 2. Preliminary finding showing the percent of records that differed in comparison to the reference standard, data element and by jurisdiction**

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Utah EHR Extract</th>
<th>Utah EBR</th>
<th>DC EHR Extract</th>
<th>DC EBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prenatal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date LMP</td>
<td>37%</td>
<td>15%</td>
<td>30%</td>
<td>70%</td>
</tr>
<tr>
<td>DateFirstPrenatalVisit</td>
<td>40%</td>
<td>13%</td>
<td>25%</td>
<td>57%</td>
</tr>
<tr>
<td>Gestational Diabetes</td>
<td>0%</td>
<td>7%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>NumberPrenatalVisits</td>
<td>77%</td>
<td>70%</td>
<td>62%</td>
<td>78%</td>
</tr>
<tr>
<td>Labor and Delivery</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antibiotics</td>
<td>18%</td>
<td>33%</td>
<td>22%</td>
<td>20%</td>
</tr>
<tr>
<td>Augmentation</td>
<td>7%</td>
<td>18%</td>
<td>7%</td>
<td>23%</td>
</tr>
<tr>
<td>Induction</td>
<td>8%</td>
<td>38%</td>
<td>7%</td>
<td>22%</td>
</tr>
<tr>
<td>FetalPresentation</td>
<td>0%</td>
<td>5%</td>
<td>7%</td>
<td>18%</td>
</tr>
<tr>
<td>RouteMethod</td>
<td>2%</td>
<td>95%</td>
<td>0%</td>
<td>60%</td>
</tr>
<tr>
<td>Newborn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NICU</td>
<td>12%</td>
<td>8%</td>
<td>8%</td>
<td>7%</td>
</tr>
<tr>
<td>ObstetricAge</td>
<td>2%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>AssistedVentilation</td>
<td>8%</td>
<td>13%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>AssistedVentilation &gt; 6 hours</td>
<td>5%</td>
<td>7%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

**Discussion**
The prenatal record is often noted as missing in operational records and when present, may be scanned into the hospital record which limits automated extraction of prenatal information. The availability of information in the EHR extract depends on where the information is documented in the EHR, and documentation practices vary by facility. Some items are not documented in the standard CCD.

**Conclusions**
Automated reporting of most of the birth certificate information is feasible using existing standards, provided that problems with the standard or facility documentation practices are addressed during implementation. In addition, ongoing assessment birth certificate information reported automatically is required to ensure quality is not sacrificed.

**References**
A Matching Algorithm between ClinicalTrials.gov and PubMed to Support the Monitoring of Gaps in Published Study Results

Adam G. Dunn, PhD¹, Florence T. Bourgeois, MD MPH²,³

¹Centre for Health Informatics, Macquarie University, Sydney, NSW, Australia;
²Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA;
³Department of Pediatrics, Harvard Medical School, Boston, MA

Introduction

Clinical trial registries such as ClinicalTrials.gov provide detailed and structured information about clinical trials, including information on the study intervention and design and whether the study has been published. This information can be leveraged for a number of functions in monitoring research activities, such as identifying and monitoring gaps in the reporting of trial results or supporting automated searching and screening processes for systematic reviews. This type of automated surveillance requires consistent and comprehensive linking of trial entries in ClinicalTrials.gov to publications in PubMed. However, there is a substantial proportion of study records on ClinicalTrials.gov that is not indexed by PubMed and the links between the registry records and the publications are missing.¹,² This problem precludes systematic monitoring of trial result reporting and analyses of reporting bias.³,⁴ It also remains as a barrier to automating searching and screening in systematic reviews,⁵,⁶ for which current natural language processing methods applied to bibliographic databases appear to be reaching a performance limit.⁷

An automated tools to produce a comprehensive matching between ClinicalTrials.gov and PubMed may improve the integration of ClinicalTrials.gov with other sources of clinical trial information.⁸ Our aim was to evaluate an algorithm for matching published reports of clinical trials with their ClinicalTrials.gov study records.

Methods

Study Data: From 68,962 interventional studies on ClinicalTrials.gov, and the 330,500 PubMed records published between January 2007 and December 2016 with publication types matching clinical studies, we selected two datasets: one for training and validation, and a separate set for testing. The first included the 12,155 registry entries with exactly one matching PubMed article published after the trial completion date. The second was a random sample of 150 registry entries completed before January 2014, and with no linked PubMed articles. For this second set, we performed a manual search in PubMed to identify articles that report results following the standard approach, limited to PubMed.⁴

Design: We used MetaMap (2014AA with word sense disambiguation) to extract medical concepts from the titles, study descriptions, and conditions in the study records in ClinicalTrials.gov, and from the titles and abstracts of the articles in PubMed. We did not match investigator or author names in the prototype because this would have required a second type of matching to be able to include it in the cross-validation described below. The similarities between registry entries and articles were defined by summing weights (defined below) for shared medical concepts.

For each registry entry we produced a ranked list of articles with the aim of ranking the matching article as high as possible. To do this, we trained the medical concept weights such that concepts found to be useful for matching study records and articles would be assigned higher weights. Weights were initialized to their Jaccard index and then iteratively refined by assessing whether increases or decreases in the weight would improve the ranking for more trials than it would degrade. The schedule for selecting concepts for this assessment was based on the frequency of their occurrence, so that more common terms were updated more often in the training procedure.

In the first experiment, we trained the concepts weights using 12,155 registry entries and articles, with 80% used in cross-validation and 20% as a holdout. In the second experiment, concept weights from the first experiment were applied to 200 registry entries and 330,500 articles. As an example of how the tool could be used to measure unreported trials, we reported the time to publication on the testing set. The time to publication curve shows the proportion of trials that remain unreported at a given time and the follow-up period for trials with no known results.

Outcome measures: The rank distribution of the correct article matches within the set of candidates for each study record; the number of articles that need to be read (NNR) per study record to find matching articles. We defined the NNR metric to be comparable to metrics for the performance of automated searching and screening methods.⁷
Results

In the 20% holdout from the training and validation set, the median rank of the correct match was 1 (IQR 1-3). For 91% (2,217/2,431) of the study records, the correct article was in the top 10 candidates. In the manual check, we found matches in PubMed for 45% (90/200) of the randomly sampled registry entries. Using the algorithm, the median rank of the correct match was 42.5 (IQR 10-255), and 27% (24/90) of the correct articles were found in the top 10. Figure 1 illustrates the NNR per study record to identify a given proportion of matching articles.

Discussion

Methods for predicting links over a bipartite graph may be useful for reducing the number of PubMed articles that need to be checked to find articles reporting results of trials on ClinicalTrials.gov. A remaining practical limitation is the need to manually search a number of articles to achieve a pre-specified level of confidence. Planned improvements include the use of other metadata such as investigator names, as well as alternative representations of the registry entries and article abstracts that improve the document similarity measures.

Figure 1. Articles identified by NNR for the (a) training holdout (2,431 articles in 2,431) and (b) testing set (90/200) articles in 33,500); and (c) estimated time to publication among 200 trials without pre-existing matches in PubMed after reading 10 (0.003% of the total), 100 (0.03%), or 1000 (0.3%) of the ranked articles per registry entry.

References

Automated Identification and Predictive Tools to Help Identify High-risk Heart Failure Patients

R. Scott Evans\textsuperscript{1,2}, Jose Benuzillo\textsuperscript{2}, Benjamin D. Horne\textsuperscript{2}, James F. Lloyd\textsuperscript{1}, Alejandra Bradshaw\textsuperscript{2}, Deborah Budge\textsuperscript{2}, Kismet D. Rasmusson\textsuperscript{2}, Colleen Roberts\textsuperscript{2}, Jason Buckway\textsuperscript{2}, Norma Geer\textsuperscript{2}, Teresa Garrett\textsuperscript{2}, Donald L. Lappé\textsuperscript{2}.

\textsuperscript{1}Medical Informatics, \textsuperscript{2}Cardiovascular Clinical Program

Intermountain Healthcare. Salt Lake City, UT

\textbf{Introduction:} Heart failure (HF) affects 2.4\% of adults in the US or nearly 6 million people. The HF cost burden was expected to reach $44.6 billion by 2015 (1) of which hospitalization represents 70\% of that cost (2). Moreover, as the prevalence of HF has increased, so has the incidence of hospital readmission and mortality (3) and HF is the leading cause for recurrent hospitalizations (4). The increasing prevalence of HF mandates a new approach to decision making and especially better communication that includes clinical decision support (CDS) to help identify high-risk HF patients as soon as possible (5). One method has been the use of predictive models to identify high-risk HF patients (6-15). However, those models are rarely used and little is known about their clinical use and their impact on patient care (6, 16). In 2010, the Intermountain Risk Score (IMRS) was published which predicted HF readmission and incident HF (17). In order to provide that information quickly to providers, a CDS application was developed to use natural language processing (NLP) to improve the identification of HF patients while another application based on the IMRS calculates their 30-day all-cause readmission risk and 30-day mortality risk. A new report is used by the Cardiovascular Clinical Program at Intermountain Healthcare each day which identifies the high risk HF inpatients that need to be closely followed using a multidisciplinary care process pathway (CPP). This presentation will report the development and use of these applications and the impact on patient care.

\textbf{Methods:} The NLP application was developed to read any type of dictated report in the electronic medical record (EMR) and use “key words/terms” along with diuretic use, B-type Natriuretic Peptide (BNP) level $> 200$ pg/mL, Ejection Fraction (EF) $\leq 40$ in the previous year, and ever been eligible to any of the Center of Medicare and Medicaid Services or Joint Commission HF core measures to identify patients with HF. The other application looks at each hospitalized HF patient each day and calculates their 30-day all-cause readmission and 30-day mortality risk scores. The risk scores were grouped and presented as low, moderate and high and based on gender. The information is then included in the Cardiovascular Clinical Program’s HF patient Identification and Risk Stratification Daily Report which was developed using Tableau Software and emailed to over 300 clinicians at 9:15am each morning. The report also includes other pertinent information to facilitate the clinicians’ work-up for each patient which comes from the Echocardiogram and laboratory data. The Cardiovascular Clinical Program initiated a new CPP for HF patients targeting those identified in the HF patient Identification and Risk Stratification Daily Report with the aim to improve the transition of HF patients from hospital to home care, decrease admissions to skilled nursing facilities and reduce 30-day readmission and 30-day mortality. A before-and-after study design was used to measure the use and impact of the new HF patient Identification and Risk Stratification Daily Report and CPP for HF patients. Patients treated using the new CPP from 02/03/2014 to 06/30/2014 were compared to patients receiving standard care at the same 354 bed hospital before the CPP from 10/01/2013 to 02/02/2014.

\textbf{Results:} The addition of NLP compared to just using ICD9 codes to help identify HF patients increased the sensitivity from 82.6\% to 95.3\% and specificity from 82.7\% to 97.5\% and a current positive predictive value of 97.45\%. The daily multidisciplinary cardiovascular planning meeting and CPP now based on the information in the HF patient Identification and Risk Stratification Report are able to identify HF patients sooner and the clinicians’ review of potential HF admissions takes only 10 vs. 40 minutes. From 02/02/2014 to 06/30/2014, 100 patients were treated using the new CPP compared to 75 patients during 10/01/2013 to 02/02/2014. The 30-day mortality rate was significantly lower (7\% vs.19\%, \(p=0.03\)) in the pathway group compared to the non-pathway group, the proportion of patients discharged to home health was significantly greater (34\% vs.19\%, \(p=0.02\)), and average length of stay was a half day less and the average hospital variable costs were $807.6 less, but not statistically significant (\(p=0.44\) and \(p=0.56\) respectively). However, there was no significant difference in the number of HF patients readmitted within 30-days in the study hospital which is the third lowest in the US for HF readmission.
Discussion: In order to improve the continuum of care for hospitalized HF patients, accurate and early identification during hospitalization is crucial. Manual identification of hospitalized HF patients is time consuming and challenged by inconsistent and poorly sensitive processes. We developed two new CDS applications and the HF patient Identification and Risk Stratification Daily report to facilitate the early identification and care of HF patients. This information is automatically provided to clinicians each day and allows them to prioritize their limited time based on the identified HF patients rather than spending time to identify them manually. Use of CDS coupled with a multidisciplinary CPP was found to be an effective method to improve HF patient identification and help to significantly reduce 30-day mortality and significantly increase patient discharges to home health rather than a skilled nursing facility. Both CDS applications and the CPP have been installed at the other Intermountain hospitals that treat HF patients and additional data and sustainability will be presented along with details of our NLP methods.

References
Automated De-Identification of Distributional Semantic Models

Gregory P. Finley, PhD1,2, Serguei V. S. Pakhomov, PhD1,3, Genevieve B. Melton, MD, PhD1,2

1Institute for Health Informatics, 2Department of Surgery, and 3College of Pharmacy
University of Minnesota, Minneapolis, MN

Introduction

Machine learning approaches to difficult problems in natural language processing (NLP) rely on abundant data for training. In the medical domain, unstructured text in electronic health record (EHR) systems provides an especially rich source of natural language data. However, EHR text contains protected health information with highly restricted access. The U.S. Health Insurance Portability and Accountability Act (HIPAA) specifies requirements for protecting confidentiality in EHR datasets used for non-clinical purposes by removing certain identifying strings, such as names and addresses.1 Performing this de-identification (de-ID) process manually, however, is prohibitively expensive for very large corpora.2 Automated methods have achieved some success,3–5 but healthcare institutions often remain hesitant to permit the release of automatically de-identified text.

We present an alternative approach: de-identifying a word co-occurrence table rather than raw text. Co-occurrence statistics lie at the heart of many distributional semantic models, which have numerous applications in biomedical NLP.6 By their nature, these models do not preserve the syntax of their source text, dramatically reducing the risk to confidentiality even before de-ID. If stripped of identifiers, these models could be safely shared with other researchers to improve outcomes in NLP and information retrieval.

The success of this approach depends upon meeting two objectives: the successful removal of identifiers from a model, and the preservation of that model’s NLP utility. We present experimental work addressing both concerns: we compare an automated de-ID strategy to manual expert de-ID of clinical notes in English, and we measure the effect of de-ID on the accuracy of an NLP tool for acronym disambiguation that relies on co-occurrence statistics.

Methods

We evaluate three variants of a simple string match de-ID algorithm: 1) any words not present in the inventory of non-proper nouns in the UMLS SPECIALIST Lexicon are removed; 2) words appearing anywhere in the names or addresses of all patients in the enterprise EHR system are removed; 3) both of the above. Numerals, months, and days of the week are removed in all conditions. Tokens are not stemmed. All string comparisons are case insensitive.

Note that many of these strings are not in themselves identifying but are rather fragments of identifiers—e.g., ‘street’, ‘from’—and that most fragments also have non-identifying uses that are far more common than their sensitive uses. With this in mind, we also test slightly relaxed variants of Condition 3 above in which the most frequently occurring words are kept in the model regardless of their status as potential identifier fragments.

To score coverage, we compare results of our de-ID method to manual expert de-ID of a corpus of clinical text (1.2 million words). Of 21,852 word types in the corpus, 3,109 can occur as identifier fragments. We measure the recall of our de-ID system over these 3,109. We also measure accuracy of an acronym disambiguation system following the automated de-ID of its co-occurrence table, scored against a manually annotated gold standard of 18,499 instances of 74 acronyms.

Results

Results in Conditions 1-3 are given in Table 1. Only Condition 3 achieves the comprehensive removal of identifiers; however, this strategy has a high false positive rate and leads to a major drop in NLP performance. (For reference, the non-de-identified acronym system achieves 90.7% accuracy.) Table 2 shows four variations of the relaxed Condition 3. Acronym accuracy increases dramatically from the unmodified Condition 3, trading with a slight impact on de-ID coverage. We found a good balance to be keeping the most common 2,000 words, which causes only a 1.3% drop in acronym model performance while still targeting nearly all potential identifiers.

We systematically examined the 44 potential identifier fragments missed by the recommended de-ID strategy to determine if they pose any risk to confidentiality. In roughly half of all cases, the string was a personal name homographic with a common word (‘will’, e.g.); in most others, the string was part of a place name (‘new’ as in
‘New York’, e.g.). These misses likely pose minimal to no risk due to the nature of co-occurrence models: uses of as an identifier are impossible to separate from innocuous uses, and because these are exceptionally common words, counts of identifier tokens are dwarfed by counts of non-identifier tokens. Of all possible identifier fragments in the 2,000 most common words, fewer than 1% of their combined occurrences in the corpus were actually in identifiers.

Discussion

We show that it is possible to create semantic models from large corpora that are nearly completely stripped of potential identifiers yet still function quite well for NLP applications. Even extremely aggressive de-ID strategies can be applied without seriously compromising the utility of distributional models; using more sophisticated and less aggressive machine learning de-ID algorithms⁴ should lessen the impact on NLP performance further.

We recommend that these methods be employed to create sharable NLP resources, which would allow researchers and institutions without access to large clinical text corpora to benefit from NLP tools developed from data available to larger medical research institutions (e.g., CTSI, PCORnet). The increased availability of such resources would serve both to expand the research community in medical NLP and to enhance information extraction and retrieval outcomes for other informatics researchers.

Acknowledgments

This work was supported by the National Institutes of Health (R01LM011364, R01GM102282 and 8UL1TR000114) and a University of Minnesota Academic Health Center Faculty Development Grant.

Table 1. De-ID recall and de-identified acronym model performance by condition.

<table>
<thead>
<tr>
<th>Cond.</th>
<th>Remove patient info?</th>
<th>Remove terms not in UMLS?</th>
<th>Identifier recall (types)</th>
<th>False pos. rate (tokens)</th>
<th>Acr. model words removed</th>
<th>Acr. model accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no</td>
<td>yes</td>
<td>.846</td>
<td>.127</td>
<td>62,077 / 88,228</td>
<td>.895</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>no</td>
<td>.886</td>
<td>.740</td>
<td>28,315 / 88,228</td>
<td>.825</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>yes</td>
<td>1.00</td>
<td>.815</td>
<td>70,441 / 88,228</td>
<td>.774</td>
</tr>
</tbody>
</table>

Table 2. Results under de-ID Condition 3 when retaining common words (our recommended approach is shaded).

<table>
<thead>
<tr>
<th># most common words retained</th>
<th>Possible identifier fragments retained</th>
<th>Identifier recall (types)</th>
<th>False pos. rate (tokens)</th>
<th>Acr. model words removed</th>
<th>Acr. model accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>common words</td>
<td>.993</td>
<td>.120</td>
<td>70,030 / 88,228</td>
<td>.873</td>
</tr>
<tr>
<td>2000</td>
<td>common words, partial cities</td>
<td>.986</td>
<td>.068</td>
<td>69,074 / 88,228</td>
<td>.894</td>
</tr>
<tr>
<td>3000</td>
<td>common personal names, cities</td>
<td>.977</td>
<td>.044</td>
<td>68,504 / 88,228</td>
<td>.897</td>
</tr>
<tr>
<td>5000</td>
<td>several names</td>
<td>.947</td>
<td>.023</td>
<td>67,414 / 88,228</td>
<td>.903</td>
</tr>
</tbody>
</table>

References

Adoption of Nurse Facing Clinical Decision Support Following EHR Implementation

Denise Goldsmith, MPH, MS, RN\textsuperscript{1}, Karen Bavuso, MSN, RN\textsuperscript{2}, Charles Lagor, MD PhD CPHIMS\textsuperscript{2}, Sarah Collins, PhD, RN\textsuperscript{1,2,3}

\textsuperscript{1}Brigham and Women’s Hospital, Boston, MA; \textsuperscript{2}Partners HealthCare, Boston, MA; \textsuperscript{3}Harvard Medical School, Boston, MA

Introduction

With the recent and rapid expansion of electronic health records in the acute hospitalization space nurses find themselves exposed to Clinical Decision Support (CDS) functionality that was previously unavailable to them. Such is the case within this Academic Medical Center (AMC) that recently implemented electronic nursing documentation, transitioning from paper. It was the objective of this study to evaluate the actions taken by nurses when presented with CDS interventions related to risk for falls, risk for skin breakdown and patient restraint use. These three domains were chosen for evaluation primarily because the CDS rules fire solely within nursing documentation and because accountability for the assessment, intervention and outcomes associated with these elements are all within the scope of nursing practice. Additionally, the elements of fall risk and skin integrity are contained within the National Database of Nursing Quality Indicators\textsuperscript{®} and appropriate management of restraint use is a standard of The Joint Commission. When presented with a CDS intervention in this electronic health record (EHR), nurses have the option to agree with and act on a CDS recommendation (accept), choose a reason for not taking the recommended action (override), or ignore it all together by not interacting with the alert and navigating out of the CDS window (ignore). The three CDS interventions we focused on are rule-based and driven off of nursing assessment documentation. Prior to CDS implementation each alert underwent a review of nursing and system workflow to determine the best time and place for the alert to fire. These CDS interventions are incorporated into the admission documentation process and are for the most part non-interruptive in that they do not require a nurse to take an action for the CDS in order to proceed with his/her documentation workflow. Minimal CDS education for nurses was included in the EHR implementation training. This podium presentation will provide an overview of all nurse-facing CDS interventions active at our institution and an in-depth evaluation of CDS interventions described above that target regulatory requirements using our most updated data sets and analyses.

Methods

CDS alert firing data for this evaluation was extracted from the EHR system used at this AMC. The data used in this analysis, extracted from a larger data set of all CDS data for the AMC, represents the alert firing activity and rates of ‘accept’, ‘override’ and ‘ignore’ for patients admitted as inpatients. Because the recommendations for these three CDS interventions were based on regulatory requirements we consider an ‘acceptance’ to be a compliant action, and an ‘override’ to be a non-compliant action, and an ‘ignore’ to be a non-compliant inaction. The three CDS interventions included in our analysis were non-interruptive alerts, presented as an interactive alert window, within the EHR admission assessment documentation screen. The CDS fired when specific assessment criteria were met (e.g., Morse Fall Scale and Braden Score above a specified threshold). The alert remained unresolved until it was either ‘accepted’ or ‘overridden’. For example, if the alert resulted in plan of care (POC) documentation (CDS recommendation was accepted), or if the nurse selected an override reason and did not complete POC documentation (CDS recommendation was overridden), the CDS would no longer be presented to the nurse. If an override reason was selected the alert was resolved for at least 24 hours or indefinitely depending on the reason selected.

Our current data set represents an eleven week time period beginning five months following EHR implementation and education of nursing staff regarding the use and value of CDS in their documentation workflow. We continuously update these data and will present our most updated data findings. For the three CDS interventions analyzed an ‘acceptance’ included the following possible actions: 1) following a link within the alert message that redirected the user to the patient’s POC or 2) placing an order as recommended by the alert. An override (“non-compliant” actions) included instances when the alert was acknowledged, commented on, or labeled insignificant by a nurse without following the link to the patient’s POC or placing an order as recommended.
After the data was extracted from the larger data set it was separated into discrete worksheets. A custom sort by Date, then Time, then Patient ID, then Follow-up action was performed. To prepare the data, we removed rows that were not a unique instance of a CDS intervention, such as “duplicate actions” that occurred for the same patient, the same nurse and the same CDS rule at the exact date and time. Once all duplicates had been removed we coded the actions as an “accepted” recommended action or an “overridden” recommended action. Each code was filtered by the type of follow-up action taken (e.g., linking to the patient’s Plan of Care (POC) or the placing of an order) to perform percent calculations.

**Results**

The three CDS interventions resulted in a total of 14,000 alerts fired over eleven weeks. Overall, nurses ignored the CDS alerts 91% to 95% of the time it was made available to them to act upon. Of the 5% to 9% of alerts that were not ignored, acceptance rates ranged from 20% to 52%. Our data indicates that overall, an alert was accepted (not ignored or overridden), only 4.6% (369/7976), 1.8% (37/2025), and 0.97% (39/3999) of the time. The Fall Risk CDS intervention which recommended that a POC problem be documented based on the patient’s Morse Fall score greater than 45 had an acceptance rate of (compliant with recommendation) 52% when the alert was not ignored. The Risk for Impaired Skin Integrity CDS, which recommended that a POC be documented, based on a patient’s Braden Score of 18 or less, had an acceptance rate (compliant with recommendation) of 32% when the alert was not ignored. The Restraint CDS recommended that a POC be documented based on the presence of an active restraint order. This CDS had an acceptance rate (compliant with recommendation) of 20% when the alert was not ignored (Table 1). The data also indicates variation in the rates of override actions among the three CDS interventions. When the Restraint alert was overridden, 23.5% of the time the nurse selected the override option indicating that the alert was ‘insignificant’. However, when the other CDS alerts were overridden, the ‘insignificant’ option was only selected 3.6% and 0.9% of the time for High Fall Risk and Risk for Impaired Skin Integrity, respectively.

<table>
<thead>
<tr>
<th>CDS</th>
<th>% CDS Ignored</th>
<th>% CDS Accepted</th>
<th>% CDS Overridden</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Fall Risk</td>
<td>91% (7265/7976)</td>
<td>52% (369/711)</td>
<td>48% (342/711)</td>
</tr>
<tr>
<td>Risk for Impaired Skin Integrity</td>
<td>94% (1910/2025)</td>
<td>32% (37/115)</td>
<td>68% (78/115)</td>
</tr>
<tr>
<td>Restraint Plan of Care</td>
<td>95% (3803/3999)</td>
<td>20% (39/196)</td>
<td>80% (157/196)</td>
</tr>
</tbody>
</table>

**Discussion**

CDS is not intended to replace nursing judgment, but rather to provide a tool to assist them in making higher quality decisions while planning the care of their patients2. For patients that met the criteria for these regulatory-based CDS interventions we observed very high “ignore” rates. Even when the CDS was not ignored, the alert was overridden by the nurse most of the time. These results generate discussion and inquiry as to whether CDS for nurses that recommend POC interventions have value within their documentation workflow. We choose these CDS to be non interruptive to the workflow. This raises the question as to whether this was the right decision. Our analysis has limitations. We based our assessment of “accepted” versus “overridden” on a simple classification scheme of follow-up actions, which may not apply in every situation. We also do not know if the nurse acted upon the CDS recommendation from outside the CDS window. In this analysis we observed variations within the data that are useful to direct refinement and optimization efforts for nursing CDS as we learn how to better support nursing decision making and add value where it is most needed.

**Conclusion**

We have begun our monitoring and analysis of nursing CDS data and will continue to update our data sets and analysis. In our preliminary work we have observed interesting variation that can direct efforts for optimization and identification of CDS interventions that are not adding value to the nurses’ workflow and decision making. Further analysis needs to be directed at why nurses choose to either ignore or take the actions they do when presented with CDS intervention recommendations.

**References**

Implementing and designing a distributed regression analysis module within PopMedNet™ for a large scale distributed data network

Qoua L. Her, PharmD, MSc¹, Jessica M. Malenfant, MPH¹, Sarah Malek, MPPA¹, Yury Vilk, PhD¹, Elizabeth Cavagnaro, MPH¹, Lingling Li¹, ScD, Jeff Brown, PhD¹, Sengwee Toh, ScD¹
¹Harvard Pilgrim Health Care Institute, Boston, MA

Introduction: To obtain sufficient sample sizes and valid regression analyses, many clinical studies require pooling of individual-level data from multiple data sources.¹ Patient privacy and the protection of propriety information have limited this practice.² However, through distributed regression analysis (DRA), statisticians have demonstrated the ability to perform regression analysis with only de-identified intermediate statistics to obtain statistically equivalent regression parameters as pooled individual-level data analysis.³, ⁴ For logistic and Cox regression models, DRA involves iterations of each data source computing site-specific intermediate statistics and transferring the statistics to an analysis center, where the statistics are aggregated and used to update parameter estimates. These parameter estimates are sent back to the data source to further fine tune the intermediate statistics and re-estimate regression parameters. This iterative process continues until the model converges.

The inability to automate DRA iterations has limited its use in practice. DRA is highly desirable in the FDA’s Sentinel system, a distributed network that uses multiple electronic healthcare data sources to monitor the safety of medical products. Ensuring patient privacy is critical for data partner collaborations in Sentinel. PopMedNet™ (PMN), the open-source distributed data-sharing platform used by Sentinel, is being enhanced to facilitate the iterative DRA process and enable users to conduct secure DRA for continuous, binary, and survival outcomes. A major development goal was to limit modifications to data partners' existing hardware and software configurations. This presentation will describe the software designs, user interfaces, workflows, and challenges of building the DRA module in PMN.

Methods: An interdisciplinary team of Sentinel investigators including epidemiologists, computer scientists, programmers, biostatisticians, and informaticians is leading the development of the DRA module. Through an iterative process, the team analyzed PMN’s existing functionalities and mapped out software designs and workflows that would automate DRA. As part of the project, the team sent a short survey to data partners to (1) inventory current hardware configurations and software heterogeneity, and (2) assess each partner’s ability and willingness to allow DRA automation. DRA module’s results will be compared to the pooled individual-level data analysis. We will conduct evaluations in a simulated distributed environment and at select Sentinel data partner sites. If the DRA module is successful, it will produce results that are statistically equivalent to the pooled analysis.

Results: A privacy-preserving automated DRA module can be developed and implemented within the existing PMN environment. The proposed DRA will involve two steps: 1) assembling an analytical dataset and structuring data fields for DRA using Sentinel’s publically available Cohort Identification and Descriptive Analytics Tool, and 2) automating the DRA iterations between data partners and the Sentinel Operations Center (SOC) with PMN’s existing distributed querying capabilities and workflow.⁵ (Figure 1). The survey identified three general hardware configurations, the existence of SAS software heterogeneity, and mixed perspectives towards full automation, in which some data partners would like to review intermediate statistics prior to transmission to the SOC.

Discussion: DRA is feasible in Sentinel’s distributed data network and will leverage existing Sentinel and PMN infrastructure. Heterogeneous configurations across Sentinel data partners’ hardware configurations and willingness to automate the process pose some challenges. Thus, the study team will
consider a semi-automated design and workflow and conduct trainings to allow full-scale adoption. A functional DRA module will still require appropriate governance to facilitate real-world application. Successful implementation of a DRA module in PMN will likely lead to greater use of DRA in multi-center studies as PMN is open-source and the informatics platform is used by several other distributed data networks, including PCORnet, National Cancer Institute’s CRNnet, and the NIH Health Care Systems Research Collaboratory.

![Diagram of the process]

**Figure 1**

**References**

Implementing Pharmacogenomic Clinical Decision Support:  
Design and Prescriber Response in the eMERGE Network

Timothy M. Herr, MS 1; Josh F. Peterson, MD, MPH2; Luke V. Rasmussen, MS1; Pedro J. Caraballo, MD3;  
The eMERGE Network

1Northwestern University Feinberg School of Medicine, Chicago, IL; 2Vanderbilt University Medical Center,  
Nashville, TN; 3Mayo Clinic, Rochester MN

Introduction
Previous literature has established that pharmacogenomic (PGx) clinical decision support (CDS) may be an ideal  
implementation avenue for genomic medicine.(1) However, analyses of PGx CDS design and resulting outcomes in  
live clinical settings have been relatively limited in scope.  To better understand design approaches and prescriber  
response to PGx CDS, we evaluated implementations and alert data at seven different sites in the Electronic Medical  
Records and Genomics (eMERGE) Network.(2)

Methods
This study was conducted among seven large academic and research-oriented health care organizations.  The scope  
of the project was limited to alerts, providers, and patients that were part of the eMERGE-PGx Project.(3)  

To understand the PGx CDS design decisions made by each site in the eMERGE Network, we surveyed key  
representatives at each participating organization.  We developed survey questions via a series of informal  
interviews conducted with one representative at each site. We then assembled a formal questionnaire and distributed  
it by email for completion by the same representatives. Sixteen questions elicited differences in design objectives  
and constraints for each site’s implementation. Key topics included the drug-gene interactions (DGI) for which PGx  
alerts were implemented (and over what time period), where in the electronic health record (EHR) alerts fired, the  
source of alert recommendations, and available actions prescribers could perform in response to alerts.  

To understand how prescribers responded to PGx CDS alerts, we asked each site to provide evidence of provider  
response based on EHR log data. As the various eMERGE sites have widely different technical infrastructures,(4)  
we developed a standard data dictionary through a series of calls conducted by the eMERGE PGx Workgroup.  This  
allowed each site to report alert data in a consistent format for aggregate analysis.  Prescriber response was divided  
into two different actions – “alert response” (corresponding to the action taken on the actual alert in the EHR) and  
“clinical response” (corresponding to the relevant clinical action taken after viewing an alert).  Alert responses were  
standardized to “Accepted,” “Overridden,” and “Ignored.” Clinical responses were standardized to “Followed,”  
“Not Followed,” and “No Action.” Each site was responsible for defining these terms for its own data according to  
local practices. Other key data points collected included alert metadata (such as the drug and recommendation),  
alert time, and prescriber department and education level. Submitted data were analyzed both in total aggregate and  
broken down according to alert characteristics.  For consistency, preliminary data analysis was limited to active,  
post-test alerts only.

Results
All seven sites implemented some form of active alerting (interruptive alerts that fire during the prescribing  
workflow), while four implemented some form of passive alerting (non-interruptive alerts visible on a general  
recommendations screen, in-basket messages, etc.). Only two sites elected to implement pre-test alerting (alerts  
suggesting prescribers order a genetic test before prescribing a particular drug). Alerts varied widely on what  
genetic data were used as inputs – some simply looked for the existence of a genotype test result of any value, others  
looked for an actionable genotype, an actionable phenotype, or a combination of both. Alert recommendations were  
typically modeled after CPIC recommendations, though with some local modifications or review. Three sites  
implemented dynamic dose calculations for warfarin prescribing, while alerts for other drugs were typically static  
recommendations. The available alert actions varied widely from site to site and alert to alert.

Preliminary analysis included 294 active, post-test PGx alerts from six sites. The seventh site had zero alerts fire for  
the DGIs of interest, though they did report recording alerts for related medications not covered in this analysis.  
DGIs examined included codeine/oxycodone/hydrocodone/tramadol (CYP2D6), clopidogrel (CYP2C19),
simvastatin \((SLCO1B1)\), and warfarin \((CYP2C9, VKORC1)\). In terms of alert response, 183 (62%) alerts were accepted, 75 (26%) were overridden, 34 (12%) were ignored, and 2 (0.8%) had an undetermined action. In terms of clinical response, 124 (42%) alerts were followed, 126 (43%) were not followed, and 44 (17%) had no relevant action taken. There was moderate to high correlation between alert response and clinical response \((r=0.697)\).

Response rates varied by DGI. Only one DGI (codeine) had a majority of alerts followed (Table). All other DGIs had a majority of alerts that were not followed.

Table. Provider Responses to Active, Post-Test PGx CDS Alerts by Drug-Gene Interaction

<table>
<thead>
<tr>
<th>Drug</th>
<th>Total Alerts</th>
<th>Alert Response</th>
<th>Clinical Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accept Override Ignore Unknown</td>
<td>Followed Not Followed No Action</td>
</tr>
<tr>
<td>Codeine</td>
<td>114</td>
<td>102 0 10 2</td>
<td>69 18 27</td>
</tr>
<tr>
<td>Clopidogrel</td>
<td>65</td>
<td>14 46 5 0 10 0</td>
<td>40 15</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>24</td>
<td>22 28 18 0</td>
<td>11 13 0</td>
</tr>
<tr>
<td>Warfarin</td>
<td>91</td>
<td>45 18 0</td>
<td>34 55 2</td>
</tr>
<tr>
<td>Total</td>
<td>294</td>
<td>183 75 34 2</td>
<td>124 126 44</td>
</tr>
</tbody>
</table>

Discussion
These results demonstrate that there is currently little consensus for how to best design and implement PGx CDS alerts, even among sites within a collaborative consortium. Though our results show encouraging indicators that PGx CDS can be effective, a majority of alerts in these first generation systems were not followed or had no clinical follow-up. Taken together, these results indicate that there is significant room for further refinement of PGx CDS systems. Additional effectiveness studies are necessary to determine why response rates were generally low. Possibilities include low relevance of alerts, incorrect timing of alerts, or unclear communication of recommendations.

This study also demonstrates the complexity of understanding alert response. Actual alert responses vary greatly from simple accept/override/ignore conventions. Not all alerts have the same available actions, and some available actions are not fully represented by the accept/override/ignore paradigm. For instance, clicking a link to review educational materials could be seen as a successfully capitalized opportunity to educate a prescriber, even if they do not follow the recommendation in that particular instance. Additionally, there is some discrepancy between the action a prescriber took on an alert and the corresponding clinical action – there is not necessarily a 1-to-1 relationship between alert and clinical response. This is likely primarily due to the design of the alerts. For instance, an alert may simply have a “cancel” function, rather than an explicit “override” function. In this case, a simple analysis of alert logging data will not reveal whether the prescriber found the alert clinically irrelevant or simply ignored it entirely. Therefore, it is necessary to analyze clinical actions to fully understand alert response.

In conclusion, these seven systems are examples of first-generation PGx CDS in live clinical settings. While these systems show promise, early analysis indicates that there is significant room for improvement. These systems were largely developed based on expert opinion, so further systematic analysis and design could potentially lead to improved relevance, usefulness, and timeliness. Such changes would likely increase alert response rates and improve patient care. Follow-up research will focus on usability and usefulness analysis.

References
Interoperability and Electronic Availability of Outside Information among US Hospitals

Authors: A Jay Holmgren,1,2 Vaishali Patel, PhD,3 Dustin Charles, MPH,3 Julia Adler-Milstein, PhD1,2

1. University of Michigan, School of Information, Ann Arbor, MI. 2. University of Michigan, School of Public Health, Ann Arbor, MI. 3. Office of the National Coordinator for Health Information Technology, Washington DC.

Introduction

Interoperability is currently a top policy priority as well as the focus of substantial effort on behalf of provider organizations. (1) (2) While national levels of electronic health record adoption have been carefully tracked over time, it has been more difficult to assess interoperability capabilities of EHRs at the national level. The Office of the National Coordinator for Health IT (ONC) has conceptualized interoperability as including four core capabilities: the ability to find information from external organizations (i.e., query), the ability to send information to external organizations, and the ability to receive information from external organizations, and the ability to integrate information received into an electronic record system without special effort on behalf of users. (3) Together, these capabilities ensure that a range of key interoperability use cases can be supported, such as emergency care where finding information is essential and transitions of care where sending and receiving information is essential. In our study, we use national data on hospital interoperability to: (1) assess the extent to which hospitals are engaging in the four core domains of interoperability; (2) identify factors associated with such engagement; and (3) identify the extent to which engagement is associated with clinical data being electronically available from outside providers.

Methods

We used data from the American Hospital Association Annual Survey for 2014 – Information Technology supplement. Our sample included 3,307 hospitals from across the United States. We supplemented this data with information from the main AHA Annual Survey, as well as publically available CMS Meaningful Use Stage 2 Attestation data.

The primary outcome measures we were interested in were the proportion of hospitals engaging in the four core domains of interoperability (finding, sending, receiving, and integrating data), the factors that correlate with hospitals engaging in all four domains of interoperability, and the factors that are associated with reported electronic availability of clinical data from outside sources when needed. We performed a multi-variate logistic regression analysis in order to identify which hospital characteristics were associated with the outcome measures of performing all four interoperability functions and having clinical data available. We used weights to produce nationally-representative results.

Results

Only 21% of US hospitals engage in all four domains of interoperability. In our multi-variate analysis of the hospital characteristics, adoption of at least a basic EHR system (4) as well as participation in a Regional Health Information Exchange Organization (5) are both strongly
positively correlated with engaging in all four domains (Odds Ratios = 1.77 and 4.46, respectively. \( P < 0.001 \)) Hospitals who use primarily one EHR vendor in their organization are also more likely to engage in all domains of interoperability (Odds Ratio = 2.65; \( P < 0.001 \)) Other hospital characteristics associated with engagement in all four interoperability domains include medium-sized hospitals (100- 500 beds) (Odds Ratio = 1.70; \( P < 0.001 \)) privately owned for-profit hospitals (Odds Ratio = 1.92; \( P = 0.003 \)) and hospitals who participate in a Patient Centered Medical Home (6) (Odds Ratio = 1.73; \( P = 0.002 \))

We found that electronically finding, receiving, and integrating data were independently positively correlated with the electronic availability of clinical data from outside sources (Odds Ratios of 4.68, 2.25, and 2.20, for each domain respectively. \( P < 0.001 \)); however, electronically sending data was not significantly associated (\( P = 0.90 \)). Hospitals that engaged in all four interoperability domains were significantly more likely to report electronic availability of clinical data from outside sources (Odds Ratio = 1.81; \( P = 0.006 \))

Discussion
While less than a quarter of US hospitals engage in all four interoperability domains, our results suggest that doing so may be key to maximizing the extent to which clinical information from outside sources is electronically available when needed. Given that certain types of hospitals were more likely to have all capabilities in place, it suggests that not all hospitals are able to pursue comprehensive interoperability or see the value in doing so. The specific characteristics that we identified point to a combination of capability reasons (e.g., those with EHRs from a single vendor may find interoperability easier) and incentive-related factors (e.g., those engaging in PCMH models may see more value from interoperability). A targeted policy strategy that simultaneously makes interoperability easier to pursue and strengthens the incentives to do may help ensure that providers have electronic access to needed information across care delivery settings.

References
5. Grossman JM, Kushner KL, November EA, LTHPOLICY PC. Creating sustainable local health information exchanges: can barriers to stakeholder participation be overcome?: Center for Studying Health System Change Washington, DC; 2008.
6. Care P-CP. Patient-Centered Medical Home. Senate Special Committee on Aging. 2007.
Does Section Order Affect Physicians’ Experiences Reviewing Ambulatory Progress Notes?

Gretchen M. Hultman, MPH¹, Jenna L. Marquard PhD², Osadebamwen Ighile MBBS, MS, Oladimeji Farri MBBS, PhD³, Elizabeth Lindemann BS⁴, Elliot Arsoniadis MD¹,⁴, Serguei Pahomov PhD¹,³, Genevieve B. Melton MD, PhD¹,⁴

¹ Institute for Health Informatics, ³ College of Pharmacy, ⁴ Department of Surgery, University of Minnesota, Minneapolis, MN
² College of Engineering, University of Massachusetts, Amherst, MA
⁵Philips Research-North America, Cambridge, MA

Introduction

Clinical notes are a vital part of the modern electronic health record (EHR). Progress notes often follow the SOAP (Subjective, Objective, Assessment and Plan) note format, established by Dr. Lawrence Weed in the 1960s as part of the Problem-Oriented Medical Record (POMR) framework (1). Clinicians report that, when reviewing a Progress note, the Assessment and Plan is the most important section that they often read first regardless of where this section appears in the note (2). Some have therefore suggested that the Assessment and Plan be located at the top of progress notes (2). Anecdotal evidence suggests that while some clinicians are writing progress notes with the Assessment and Plan section at the top, some continue to follow the SOAP format with the Assessment and Plan at the end, others use customized note templates, resulting in progress notes with variable sequences of sections. The purpose of this study was to determine whether the order in which sections appear in progress notes affects how clinicians’ review and synthesize patient cases.

Methods

This IRB-approved study was conducted at a large Midwestern Academic Health Center. A previously-described EHR system prototype (4) designed to look like CPRS/VistA was populated with four de-identified patient cases, each with nine progress notes (4). Cases were designed to represent realistic patients and to be of relatively similar difficulty. Each case contained progress notes that could be presented in one of four different section orders: SOAP (Subjective, Objective, Assessment and Plan), APSO (Assessment and Plan, Subjective, Objective), SAPO (Subjective, Assessment and Plan, Objective) and Mixed (the section orders were inconsistent across, with three notes of each of the above orders). A convenience sample of participants (n=23), internal medicine and surgery residents, was recruited for this study. Sample size was restricted by ability to recruit residents. Participants were seated at a desktop computer with the EHR opened to the notes section of the first patient case. Participants were asked to review the existing notes for that patient as they normally would and provide a verbal summary of the case. Upon completing the verbal summary, participants rated their perceived workload for the case using the NASA-TLX instrument (5). Participants repeated this process three more times, for a total of four patient cases. Participants were shown the cases in the same order. Each case had a different note order, with the note orders randomized across participants using a Latin squares design. At the end of the data collection session, participants completed an exit questionnaire that contained demographic questions as well as Likert-type questions in which participants self-reported their experience level with EHRs in general and with different specific commercial EHRs.

Results

Table 1 shows participant characteristics, including their perceived prior experience using two EHR systems in clinical practice. Most considered themselves “average” users of the EHRs. The results in Table 2 show that there was a borderline significant difference in time spent reading the notes (p=0.048) across the orders, with participants taking the shortest amount of time to read APSO ordered notes and the longest amount of time to read Mixed ordered notes. Post-hoc comparisons of all groups were preformed using a T-test and a Bonferroni adjustment for multiple comparisons. The difference in mean time spent reading was statistically significant between the APSO order and the Mixed order (p=.021). All other comparisons between mean reading times were not statistically significant. There was no significant difference in time spent on the verbal summaries (p=0.173) across the orders. There was no difference in average NASA-TLX perceived workload scores (0-50 scale) across the case orders (p = 0.650) (Table 3). Mean reading time was 10.0 minutes for surgery residents and 12.3 minutes for medicine residents. Mean
summarizing time was 1.6 minutes for surgery residents and 2.4 minutes for medicine residents. These differences were statistically significant for both reading time (p=.01) and summarizing time (p=.01).

Discussion
While participants perceived notes orders to be of equal workload, there was a significant difference in time spent reading the notes. Participants read the APSO ordered notes fastest; perhaps because they found the information they wanted more quickly. Participants read the Mixed Order notes slowest, suggesting that customized templates and the lack of an expected order for progress note sections may have negative consequences for note readers. The average reading time difference between APSO and Mixed Orders was two minutes which, scaled across many patients, is substantial. While there was no time difference for the verbal summaries, we did not examine whether the content and quality of those summaries differed in the current pilot study. Difference between medicine residents and surgery residents indicate that the two groups reviewed and summarized patient cases differently and future research should examine these differences further.

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

Table 1: Participant Demographics

<table>
<thead>
<tr>
<th>Gender</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Male</td>
<td>14</td>
</tr>
<tr>
<td>• Female</td>
<td>9</td>
</tr>
<tr>
<td>Mean Age (SD)</td>
<td>29.9 (2.48)</td>
</tr>
<tr>
<td>Mean Years since graduation (SD)</td>
<td>2.82 (1.47)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Department</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Internal Medicine</td>
<td>15</td>
</tr>
<tr>
<td>• Surgery</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CPRS/VistA experience (n=22)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• No experience</td>
<td>2</td>
</tr>
<tr>
<td>• Novice user</td>
<td>1</td>
</tr>
<tr>
<td>• Average user</td>
<td>17</td>
</tr>
<tr>
<td>• Expert user</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Average Times, by Note Order

<table>
<thead>
<tr>
<th>Note Order</th>
<th>Reading (min) Mean (SD)</th>
<th>Verbal Summary (min) Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOAP</td>
<td>11.6 (4.95)</td>
<td>2.1 (1.66)</td>
</tr>
<tr>
<td>APSO</td>
<td>10.6 (1.85)</td>
<td>1.9 (1.49)</td>
</tr>
<tr>
<td>SAPO</td>
<td>11.3 (2.29)</td>
<td>2.3 (1.62)</td>
</tr>
<tr>
<td>Mix</td>
<td>12.5 (2.12)</td>
<td>2.1 (1.04)</td>
</tr>
<tr>
<td>Average</td>
<td>11.5 (2.08)</td>
<td>2.1 (1.46)</td>
</tr>
</tbody>
</table>

Table 3: Average Workload Score, by Note Order

<table>
<thead>
<tr>
<th>Note Order</th>
<th>Average NASA-TLX Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOAP</td>
<td>30.6 (10.57)</td>
</tr>
<tr>
<td>APSO</td>
<td>31.3 (8.75)</td>
</tr>
<tr>
<td>SAPO</td>
<td>31.9 (7.04)</td>
</tr>
<tr>
<td>Mix</td>
<td>31.7 (7.78)</td>
</tr>
<tr>
<td>Average</td>
<td>31.4 (8.52)</td>
</tr>
</tbody>
</table>

Acknowledgements
The Agency for Healthcare Quality and Research (R01HS022085) and the National Science Foundation (1150057) supported this work. The content is solely the responsibility of the authors and does not represent the official views of AHRQ.
Using Social Networking Analysis to Provide Greater Context to the Evaluation of a Team-Based Communication Tool

Trevor Jamieson, MD MBI1 2 3 4, Amna Husain, MD MPH1 5, Peter Weinstein, BEng MHSc6, Teja Voruganti, MD/PhD (c)7, Allison Kurahashi, BSc Med5, Alyssa Bertram, BA5, Michael Chiang, MD MA4

1University of Toronto, Toronto, Ontario; 2St Michael’s Hospital, Toronto, Ontario; 3Women’s College Hospital Institute for Health System Solutions and Virtual Care (WIHV), Toronto, Ontario; 4Oregon Health & Science University, Portland, Oregon; 5Temmy Latner Centre for Palliative Care, Mount Sinai Hospital, Toronto, Ontario; 6Centre for Global eHealth Innovation, University Health Network, Toronto, Ontario; 7Institute for Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario

Abstract

Social Networking Analysis (SNA) is a discipline rooted in graph theory, sociology and anthropology allowing one to mathematically analyze the relationships between actors in a network in order to assess how the network functions and the various roles that actors play. We are exploring the use of SNA to more broadly define contextual factors that impact the success of LOOP, a cross-institutional social networking application centered on the patient. Audit-trail data from the intervention arm of an ongoing randomized controlled trial of LOOP, comprising messaging data from 12 collaborative teams, were used to show how SNA can be used to generate further hypotheses and refine qualitative analyses of social networking applications in general.

Introduction

Healthcare has been variously described as a complex adaptive system, a system that is non-deterministic and where conventional quantitative evaluations may fail to adequately provide the context required to understand, refine and generalize tools and solutions.1 A greater understanding of social networks is considered a window into the aggregate complexity of the system,2 and social networking analysis (SNA), an analytical discipline rooted in graph theory, sociology and anthropology that allows for rigorous understanding of those networks and the roles of actors within them,3 has been advocated by the Agency for Healthcare Research and Quality (AHRQ) as a way to understand the behavior of healthcare coordination networks specifically.4 Research on innovative collaborative teams across industries suggests that successful teams have high density of interaction and connectedness and low degrees of strict hierarchy and ‘core-periphery’ structures.5,6

LOOP is a social networking application with patients and caregivers as core members of distributed online teams seeking to improve collaboration and patient-centered outcomes;7 LOOP is currently in the midst of a pragmatic randomized controlled trial (RCT) in patients with advanced malignancy. Messaging in LOOP is open within a secure patient-oriented group; there is no private directed messaging. We applied SNA as an adjunct to the evaluation of LOOP in order to provide greater contextual understanding of the impact of LOOP on patients, caregivers and healthcare providers.

Methods

We used audit-trail data from 12 teams in the intervention arm of the LOOP RCT to quantify the level of engagement within teams. A directed tie between nodes (team members) X and Y is defined when any of the following occur: X marks a message as “attention to” Y, X replies to Y’s message, or X mentions Y in a message. SNA metrics (density, Krackhardt’s measures of hierarchy,8 and correlation with idealized core-periphery models) were calculated for the 12 teams using the UCINET software suite.9

Teams were ranked into quartiles in regards to their performance on each metric. The percentage of time a team was in the top quartile across metrics was correlated with a simpler metric of successful collaboration, the total number of messages sent within the team. If a straight count of messages sent in a team is an appropriate metric of collaboration, we would expect a strong correlation between this metric and the frequency with which a care team performs highly on the SNA metrics.

Results

Three teams (Teams 2, 3 and 11) had no activity. The graphical depictions (sociograms) and team compositions of the 9 teams with activity are shown in Figure 1. Of the teams with no activity, Team 2 was comprised of one patient and three physicians, Team 3 of one patient, one caregiver and two physicians and Team 11 of one patient and one physician.

LUB in the 9 teams with activity was universally 1.0, providing no useful information. One team, Team 8, was in the top quartile 100% of the time but was tied for 5th in terms of number of messages sent; another team, Team 1, sent the most messages but was in the top quartile on SNA metrics 0% of the time (Table 1). Overall, there was no correlation between the...
number of messages sent in a team and the percentage of time a team was found in the top quartile across the five metrics ($r = 0.30$). The sociograms of Teams 1 and 8 visually show the variation in collaboration across the two teams (Figure 1).

![Sociograms of Teams 1 and 8](image)

Figure 1. Sociograms of engagement in the 9 teams with activity. Team number in red. Node size proportional to outdegree. Green = patient/caregiver, Blue = physician, Purple = allied health provider, Orange = care coordinator, Grey = LOOP administrator.

Table 1. Number of messages sent and values for various SNA metrics across the 9 teams with activity. For each SNA metric (density, connectedness, hierarchy, efficiency and % in core), the values within the top quartile on that metric are highlighted in yellow. The final column shows the percentage of time a team’s SNA metrics were in the top quartile across the 5 metrics.

<table>
<thead>
<tr>
<th>Team</th>
<th>Messages Sent</th>
<th>Density</th>
<th>Connectedness</th>
<th>Hierarchy</th>
<th>Efficiency</th>
<th>% in Core</th>
<th>% Top Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>0.13</td>
<td>0.40</td>
<td>0.80</td>
<td>1.00</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.50</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>67</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.20</td>
<td>0.30</td>
<td>0.80</td>
<td>1.00</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>0.35</td>
<td>0.60</td>
<td>0.00</td>
<td>0.67</td>
<td>80</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>0.47</td>
<td>1.00</td>
<td>0.33</td>
<td>0.70</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0.90</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>0.40</td>
<td>1.00</td>
<td>0.75</td>
<td>0.50</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>0.33</td>
<td>0.50</td>
<td>0.67</td>
<td>0.00</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>0.67</td>
<td>1.00</td>
<td>0.50</td>
<td>0.00</td>
<td>25</td>
<td>60</td>
</tr>
</tbody>
</table>

Conclusion

SNA can provide additional information beyond more simplistic metrics of collaboration, even in situations with relatively sparse data, and is a useful adjunct to the evaluation of social networking interventions. SNA uncovers teams that are strongly collaborative but who might be passed over as entirely average by a ‘messages sent’ metric (Team 8) as well as teams who, despite higher levels of messaging activity, seem to be collaborating poorly (Team 1). This provides the opportunity for enriched contextual understanding of tools through further targeted analysis of higher and lower performing teams.

References

Quantifying the Effect of Data Quality on the Correctness of an eMeasure

Steven G. Johnson, PhD1; Stuart Speedie, PhD, FACMI1; Gyorgy Simon, PhD1; Vipin Kumar, PhD2; Bonnie L. Westra, PhD, RN, FAAN, FACMI1,3

1University of Minnesota, Institute for Health Informatics; 2University of Minnesota, Department of Computer Science; 3University of Minnesota, School of Nursing

Introduction
The increasing secondary use of electronic health record (EHR) data to improve health outcomes is promising, but it depends on clinical information being of sufficiently high quality to support the secondary use.1 eMeasures, which are quality metrics derived from EHRs, are standardized performance measures that quantify how well patient care is meeting best practices. Correctly computing an eMeasure depends on how well the data are recorded in the EHR. Data may be adequate to document care but insufficient to support the computation of an eMeasure.2 Data may be missing, incorrect, out of range or inappropriate for the field. In these situations, the patient’s record may be excluded from the calculation of the eMeasure resulting in biased results as the eMeasure could not be applied to the entirety of the population it was intended to assess. Secondary uses of EHR data can be better trusted if the impact of the underlying data quality is assessed.3 An eMeasure was used in this research as an example secondary use, but the method described in this paper can be generalized to other secondary uses.

Recent work to define data quality concepts as an ontology improves the ability to discuss data quality issues and leads to an assessment method that allows data sets to be characterized along a number of data quality dimensions.4 Terms that are capitalized and in “CamelCase” refer to the specific aspects of data quality defined in that ontology. The purpose of this study was to quantify the impact of two data quality issues, data missingness (RepresentationCompleteness) and data conformance to a domain model (DomainConstraints), on the correctness of an eMeasure (CMS178). Data quality issues were artificially created by systematically degrading the underlying quality of sample EHR data using two methods: independent and correlated degradation. A linear model that describes the change in the correctness of the eMeasure was developed to quantify the impact of data quality issues for each domain concept (DomainConcept) on the eMeasure.

Methods
After IRB approval, a de-identified random sample of 72,127 patient encounters was obtained from a clinical data repository (CDR). The CMS178 eMeasure was used as an example secondary use to illustrate the assessment process. The definition of CMS178 is “Urinary catheter removed on Postoperative Day 1 (POD 1) or Postoperative Day 2 (POD 2) with day of surgery being day zero.”5 CMS178 is a ratio where the denominator includes all hospital patients (age 18 and older) that had surgery during the measurement period with a catheter in place postoperatively. The denominator exclusions are 1) patients who expired perioperatively, or 2) patients who had physician documentation of a reason for not removing the urinary catheter postoperatively, or 3) patients who had medications administered within 2 days of surgery that were diuretics, IV positive inotropic and vasopressor agents or paralytic agents. The numerator is the number of patients in the denominator whose urinary catheter was removed within 48 hours.

A data quality assessment process was previously developed to compute RepresentationCompleteness (data missingness) and DomainConstraints (how well data conform to the domain model).4 The current research extends that approach by quantifying the degree that data quality issues for each DomainConcept impact a task. This was accomplished by deliberately changing the underlying EHR data in a systematic way and observing how those changes affected the calculated value of the eMeasure. The unmodified sample CDR data was the baseline to compare against for correctness. A variable, missing events, was computed that quantifies the correctness of the CMS178 eMeasure after the data are modified. This variable represents the number of patients that had a catheter removed within 48 hours in the baseline data but after the data set was degraded, were subsequently not counted as satisfying the CMS178 criteria. These are events of interest that were missing due to the data quality issues. The full degradation process consists of iteratively applying the degradation method to the data for each of the DomainConcepts in the domain model. The task was performed (in this case, computing CMS178 and missing events) and RepresentationComplete and DomainConstraint were recomputed on the degraded data for every DomainConcept. These results, the CMS178 eMeasure, and missing events were recorded in an analysis database. Two approaches to degrading data were examined: 1) independent and 2) correlated.

To quantify the effect on the eMeasure of degrading each DomainConcept, linear regression modeling was utilized. A model was fit to missing events as the dependent variable, with the data quality results for
RepresentacionComplete and DomainConstraints for each DomainConcept as the predictor variables. Negative changes (degradation) to the data increase missing events and can be used to quantify what would happen if instead, data quality improved. If data in an EHR is of low quality (i.e. the degraded data) and a method existed to somehow improve it by correcting the data (assuming the incorrect data could be identified) then missing events would be reduced.

Results
The results of the linear model for RepresentationCompleteness showed that the birth date, admission type, medication start date, catheter duration and catheter rationale for continued use were significant and had an impact on missing events. The coefficients can be interpreted to quantify the magnitude of the impact on missing events for a 1 unit improvement in an aspect of data quality. The model showed that for every 1% reduction in RepresentationComplete data quality for admission type, there were approximately 1% more events missed and the eMeasure is further from its correct value. But a 1% reduction in data quality for catheter rationale for continued use only results in 0.11% cases being missed. Degrading the domain constraints yielded a different set of variables that were impactful compared to RepresentationComplete. For independent degradation, the DomainConcepts of death date, medication start date and medication end date were significant. But only medication end date had an appreciable impact. A 1% improvement in data quality for medication end date reduces the missed cases by 0.38%, but the other DomainConcepts only impact the missed events by less than 0.1%.

Discussion
The results of this study show that: 1) data quality issues for different variables impact the correctness of an eMeasure and the impact can be quantified and 2) the CMS178 eMeasure, as currently defined, depends on how data issues are handled, which may misrepresent how well an organization is meeting the best practice goal of removing catheters within 48 hours of surgery. Degrading each DomainConcept independently compared to degrading in a correlated manner produced approximately the same set of variables that were most impactful. Degrading the domain constraints had a smaller impact than missing data. This result is likely because the data is only altered to violate the constraints, it is not missing, so it can still be used for the computation of CMS178.

The second finding is that the CMS178 eMeasure may not adequately measure catheter removal within 48 hours of surgery. Decisions for how to handle data that are missing or that violate domain constraints can influence how well CMS178 measures catheterization best practices. As the data set is changed, one way to handle the data quality issues is to remove affected patient encounters from both the denominator and numerator. Even though missing events increase as the underlying data is degraded, CMS178 itself does not appreciably change because it is a proportion. Handling the affected patient encounters in a different way, however, could change CMS178, which highlights a potential problem with the definition of CMS178, and eMeasures in general. They should define explicit rules for handling missing and constraint violating variables and require data quality metrics be reported with the eMeasure.

One limitation of this research is that only two approaches to injecting errors into the data were explored but there may be other types of data quality issues that occur in practice. Also, the approach to data missingness does not take into account different kinds of missing data (i.e. missing due to omission). The usefulness of characterizing data quality using these methods enables healthcare organizations to prioritize data quality improvement efforts to focus on the areas that will have the most impact on correctness and assess whether the eMeasure values that are being reported should be trusted.

References
The Spectrum of Insomnia-Associated Comorbidities in an Electronic Medical Records Cohort / AMIA 2016 Annual Symposium, November 12 - 16, 2016, Chicago, IL.

Uri Kartoun PhD1,2,*, Andrew L Beam PhD2,3,*, Jennifer K Pai ScD MHS4, Arnaub K Chatterjee MHA MPA4, Timothy P Fitzgerald PhD4, Isaac S Kohane MD PhD2,3,*, Stanley Y Shaw MD PhD1,2,*

1. Center for Systems Biology; Center for Assessment Technology & Continuous Health (CATCH), Massachusetts General Hospital, Boston, MA.
2. Harvard Medical School, Boston, MA.
3. Department of Biomedical Informatics, Harvard Medical School, Boston MA.
4. Merck & Co., Inc.

* Contributed equally.

**Objective:** Identify comorbidities enriched in patients with the diagnosis of insomnia, by interrogating an electronic medical records (EMR) database.

**Materials and Methods:** We studied individuals with diagnosis codes or medication prescriptions for insomnia in a population of 314,292 patients from two urban tertiary-care hospitals between 1992 and 2010. We extracted structured EMR variables (e.g., demographics, billing codes, and medications) and unstructured variables related to insomnia or comorbidities from narrative notes. We developed a case-control methodology to match insomnia patients to non-insomnia controls, and calculated the enrichment of comorbidities specifically in insomnia patients. For each case with insomnia diagnosis codes, we applied 1:1 matching to identify a control (non-insomnia) patient with no insomnia diagnosis codes. Controls were matched for gender and age because all patients in our database had properly documented genders and dates of birth. Additional matching criteria included the total number of facts in the EMR associated with each case or control (including the number of laboratory measurements, prescriptions, diagnosis / procedure codes, and notes). Our rationale was that patients with similar numbers of medical facts are likely to utilize health-care resources equivalently. In addition to the absence of insomnia diagnosis codes, we achieved further confirmation by selecting controls only if no sleep medications were found in the patients’ medical profiles. Only patients 18 years of age or older were eligible to be selected as cases or controls. We calculated the case-to-control enrichment (i.e., insomnia-to-non-insomnia ratio) by dividing the ascertained values for insomnia cases by the corresponding values for non-insomnia controls in the 12 months prior to the first diagnosis code or medication prescription for insomnia. The case-to-control ratio was calculated for each comorbidity and reflected the enrichment level of the comorbidity in insomnia relative to non-insomnia controls. We compared categorical variables using the chi-square test. Differences in means of continuous variables were compared using the t-test or Wilcoxon rank sum test, as appropriate. All statistical tests were 2-sided tests with Bonferroni correction for multiple comparisons of 59 covariates. We also assessed clinical variables associated with insomnia by penalized logistic regression and we used the bootstrap procedure to calculate confidence intervals. We further separately examined enrichment of comorbidities in insomnia patients in the inpatient versus outpatient setting, to better understand the potentially different practice patterns.

**Results:** In patients with insomnia-related diagnosis codes or medications, concepts related to insomnia were highly enriched in narrative notes. We find highly significant enrichment of several comorbidities in insomnia patients, including all 10 of the conditions that contribute to patients with “multiple chronic conditions”. The top-ranked comorbidities by logistic regression were also highly ranked in our enrichment analysis. Narrative mentions of insomnia-related concepts were enriched in notes from outpatient but not inpatient encounters.

**Conclusion:** Our results highlight the importance of analyzing narrative notes to understand the scope of conditions such as insomnia that are challenging to study using structured variables alone. By systematically identifying common comorbidities that co-exist with insomnia, this report can clarify the medical impact of insomnia.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>p-value</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep Disorder</td>
<td>U</td>
<td>6.5·10^{-12}</td>
<td>12.1</td>
</tr>
<tr>
<td>Alzheimer’s</td>
<td>S</td>
<td>3.6·10^{-11}</td>
<td>9.6</td>
</tr>
<tr>
<td>Sleep Apnea</td>
<td>S</td>
<td>4.3·10^{-10}</td>
<td>6.0</td>
</tr>
<tr>
<td>Anxiety or Depression</td>
<td>S</td>
<td>6.1·10^{-215}</td>
<td>4.7</td>
</tr>
<tr>
<td>Alcohol Use</td>
<td>U</td>
<td>2.4·10^{-34}</td>
<td>4.7</td>
</tr>
<tr>
<td>Alzheimer’s</td>
<td>U</td>
<td>3.2·10^{-49}</td>
<td>4.0</td>
</tr>
<tr>
<td>CHF</td>
<td>S</td>
<td>8.8·10^{-47}</td>
<td>4.0</td>
</tr>
<tr>
<td>Psychiatric Disorder</td>
<td>S</td>
<td>6.4·10^{-45}</td>
<td>3.9</td>
</tr>
<tr>
<td>Sleep Apnea</td>
<td>U</td>
<td>1.0·10^{-42}</td>
<td>3.8</td>
</tr>
<tr>
<td>CHF</td>
<td>U</td>
<td>1.1·10^{-38}</td>
<td>3.8</td>
</tr>
<tr>
<td>Anxiety or Depression</td>
<td>U</td>
<td>&lt;1.0·10^{-609}</td>
<td>3.7</td>
</tr>
<tr>
<td>Psychiatric Disorder</td>
<td>U</td>
<td>3.1·10^{-79}</td>
<td>3.7</td>
</tr>
<tr>
<td>Non-Viral Hepatitis</td>
<td>S</td>
<td>1.3·10^{-16}</td>
<td>3.6</td>
</tr>
<tr>
<td>COPD</td>
<td>S</td>
<td>5.3·10^{-48}</td>
<td>3.5</td>
</tr>
<tr>
<td>NAFLD</td>
<td>S</td>
<td>1.5·10^{-14}</td>
<td>3.3</td>
</tr>
<tr>
<td>Gastrointestinal Disorder</td>
<td>U</td>
<td>8.0·10^{-145}</td>
<td>3.3</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>U</td>
<td>6.3·10^{-45}</td>
<td>3.3</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>U</td>
<td>1.4·10^{-12}</td>
<td>3.1</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>S</td>
<td>2.7·10^{-9}</td>
<td>3.1</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>S</td>
<td>7.0·10^{-41}</td>
<td>3.0</td>
</tr>
<tr>
<td>Viral Hepatitis</td>
<td>S</td>
<td>1.2·10^{-7}</td>
<td>2.9</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>U</td>
<td>1.2·10^{-72}</td>
<td>2.9</td>
</tr>
<tr>
<td>Atrial Fibrillation / Atrial Flutter</td>
<td>S</td>
<td>4.2·10^{-34}</td>
<td>2.7</td>
</tr>
<tr>
<td>Asthma</td>
<td>U</td>
<td>3.3·10^{-48}</td>
<td>2.7</td>
</tr>
<tr>
<td>Asthma</td>
<td>S</td>
<td>2.1·10^{-37}</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Figure 1. Enrichment of comorbidities in 7,604 insomnia patients and 7,144 matched non-insomnia patients, S = Structured; U = Unstructured. The values used to calculate enrichment in insomnia reflect the average count per patient of structured or unstructured variables in the 12 months prior to the first insomnia diagnosis code.

Figure 2. Enrichment of top 10 comorbidities in insomnia patients: inpatient versus outpatient, S = Structured; U = Unstructured. A. Inpatients only. B. Outpatients only.
Findings of an Online Delphi Consensus Panel for Stakeholder-determined Person-centered Outcomes Research Priorities

Katherine K. Kim, PhD, MPH, MBA1, Dmitry Khodyakov, PhD2, Kate Marie, MPA1, Marika Booth2, Paul A. Heidenreich, MD3, Zhaoping Li, MD4, Michael K. Ong, MD4, Jane Burns, MD6, Daniella Meeker, PhD5, Lucila Ohno-Machado, PhD, MD, MBA6
1University of California Davis, Sacramento, CA; 2RAND, Santa Monica, CA; 3Stanford, Palo Alto, CA; 4University of California Los Angeles, VA Greater Los Angeles Health Care System, Los Angeles, CA; 5University of Southern California, Los Angeles, CA; 6University of California San Diego, San Diego, CA.

Introduction

Engagement of patients in research is critical to generating evidence that is relevant to their concerns and values. There is meager understanding of optimal engagement in early phases of research, such as the selection of relevant topics for person-centered outcomes research (PCOR). High-touch methods such as in-person focus groups and interviews are not easily scaled to larger populations, while high-tech methods such as online surveys and conference calls suffer from superficial interaction. This study addresses this gap by assessing deep, high-tech engagement with stakeholders of pSCANNER: Patient-centered SCAlable National Network for Effectiveness Research, a clinical data research network of 21 million patients.

Methods

We conducted a series of stakeholder panels for weight management/obesity (WMO), heart failure (HF), and Kawasaki disease (KD) to set PCOR priorities, using online software and a well-documented modified-Delphi method, a deliberative and iterative approach to attaining consensus with discussion and statistical feedback. The topics were rated using five criteria—informative decision making, collaboration, relevance, impact, innovation—applied to aspirational goals: i.e., to decrease the proportion of adults who are obese by 10% by 2020, to reduce unwarranted hospital readmissions for HF by 25% by 2020, and to reduce the percentage of new Kawasaki disease patients with permanent disability by one-quarter (from 7% to 5.25%) by 2020. Panelists were adults with overweight/obesity, HF, or Kawasaki disease (including parents for Kawasaki disease), clinicians and researchers recruited nationally via social media or pSCANNER clinical sites. Within condition, participants were randomized to either a single-stakeholder (patient or clinician) or mixed-stakeholder (patients, clinicians, and researchers) blinded panel. Only two panels were fielded for KD due to overlap among clinicians and researchers. Participation was online, asynchronous, and took four hours over four-weeks. Analysis included rank-ordered topics (based on mean score), determination of consensus (positive agreement) within panels on each criterion, and comparisons across panels.

Results

349 panelists participated in the project. Retention rate, the percent of panelists who participated in both round 1 and round 3 voting, was 67% for WMO, 68% for HF, and 97% for KD. (Table 1)

Table 1. Expert Panel Participants

<table>
<thead>
<tr>
<th>Weight Management/Obesity (n=121)</th>
<th>Heart Failure (n=112)</th>
<th>Kawasaki Disease (n=116)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients only (n=32)</td>
<td>Patients only (n=30)</td>
<td>Patients only (n=50)</td>
</tr>
<tr>
<td>Clinicians only (n=38)</td>
<td>Clinicians only (n=30)</td>
<td></td>
</tr>
<tr>
<td>Mixed: Patients, clinicians, researchers (n=51)</td>
<td>Mixed: Patients, clinicians, researchers (n=52)</td>
<td>Mixed: Patients, clinicians, researchers (n=66)</td>
</tr>
</tbody>
</table>
In WMO, the panels altogether achieved consensus on three of nine topics. HF panels did not achieve full consensus on any of the seven topics and KD panels achieved consensus on two of seven topics. (Table 2) The top research priorities in WMO and HF by rank (by ratings of criteria on a 1-9 scale) and overall mean score (averaged across all panels) were related to care coordination and lifestyle self-management while medications and surgery were ranked lowest. In contrast, the top priorities in KD were related to medications and diagnostic tests while the lowest were lifestyle related. In WMO, clinicians’ ratings differed significantly from patients and the mixed panel. HF patients differed from clinicians and the mixed panel. KD showed the least disagreement between patient and mixed panels.

Table 2. Top Ranked PCOR Priorities by Panel and Overall Mean Score

<table>
<thead>
<tr>
<th>Topic</th>
<th>Rank by</th>
<th>Clinician</th>
<th>Patient</th>
<th>Mixed</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMO: Compare different lifestyle changes such as nutrition, physical activity, mindfulness, and stress reduction (alone or in combinations), to determine which ones are most effective in maximizing weight loss or maintaining weight loss for individuals who are obese.</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>6.98</td>
<td></td>
</tr>
<tr>
<td>WMO: Compare different lifestyle change support strategies such as health coaching, individually tailored health messages, or communications with clinicians (alone or in combinations), to determine their effect in maximizing weight loss and/or maintaining weight loss for individuals who are obese.</td>
<td>2</td>
<td>2</td>
<td>6.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMO: Compare the effectiveness of support by peer groups to support by healthcare teams on individuals' self-management capacity and their weight loss.</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6.87</td>
<td></td>
</tr>
<tr>
<td>HF: Understand whether and how care coordination activities impact patient outcomes.</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7.33</td>
<td></td>
</tr>
<tr>
<td>HF: Understand the balance between patient-centered outcomes and the patient effort required to achieve them.</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>7.24</td>
<td></td>
</tr>
<tr>
<td>HF: Compare the effectiveness of remote monitoring and telehealth strategies for heart failure.</td>
<td>1</td>
<td>7.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD: Compare medications for infants and children with specific characteristics to determine which ones have the best long term outcomes with the least risk.</td>
<td>2</td>
<td>1</td>
<td>7.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD: Compare the effectiveness of diagnostic tests for early detection of KD.</td>
<td>1</td>
<td>2</td>
<td>7.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD: Understand whether and how algorithms can be applied to electronic health record data (e.g., patterns of symptoms or test results) to identify possible cases of KD.</td>
<td>3</td>
<td>7.14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion and Conclusion

We demonstrated that structured, high-tech engagement of large numbers of patients and other stakeholders in prioritization using online tools can yield rich contributions for PCOR. In this first of its kind study of substantial-scale stakeholder engagement in PCOR prioritization, results differed by health condition and stakeholder group. We plan to conduct analysis of online discussions which may offer insights as to why. Our results raise further research questions on how to combine and balance the input of diverse groups, particularly patients and caregivers.

References

Developing a Harmonization Process between a Nursing Reference Terminology and a Classification System

Tae Youn Kim, PhD, RN1,2, Nicholas Hardiker, PhD, RN, FACMI2,3, Kay Jansen, DNP, PMHCNS-BC, RN2,4, Amy Coenen, PhD, RN, FAAN2,4

1University of California Davis, Sacramento, CA, USA
2International Council of Nurses, Geneva, Switzerland
3University of Salford, Salford, UK
4University of Wisconsin-Milwaukee, Milwaukee, USA

Introduction

Health and nursing terminologies keep evolving to incorporate advances in science, changes in practice, and health policy. Harmonizing these terminologies is of importance to enhance the coverage of domain knowledge of interest and ensure interoperability of clinical data across multiple disciplines and settings. Of various efforts and initiatives established to cross-map diverse terminologies, harmonization of International Classification for Nursing Practice (ICNP®) and the Clinical Care Classification (CCC) is under way through a formal collaborative agreement between International Council of Nurses (ICN) and SabaCare, Inc.1 This collaborative work promotes bi-directional transformations of electronic health information and complements each system’s strength in the eHealth practice environment. The purpose of this study was to a) examine differences in structure and conceptual granularity between ICNP and CCC, and b) establish a harmonization process between a reference terminology system (i.e., ICNP®) and a classification system (i.e., CCC).3 It is expected that the outcome of this study will facilitate on-going collaborative efforts as both systems evolve.

Methods

Using ICNP nursing diagnoses (n=720) and intervention (n=735) concepts, this study involved a 3-step process of harmonizing ICNP and CCC: 1) identifying conceptual equivalency between ICNP and CCC, 2) determining conceptual similarity (i.e., ICNP concepts are narrower or broader in meanings) between ICNP and CCC, and 3) clustering unmapped ICNP concepts using the highest level concepts in CCC (i.e., 21 care components which provide a standardized framework for organizing nursing diagnoses and interventions). Conceptual equivalency was determined based on pre-existing equivalencies4 within the Unified Medical Language System (UMLS) maintained by the U.S. National Library Medicine5, followed by experts’ reviews of the candidate equivalencies. Of the ICNP concepts that were not mapped to CCC, conceptual similarity was sought manually by reviewing the formal computable definition of each ICNP concept modeled using Web Ontology Language (OWL) in Protégé environment2 and the definition of the remaining CCC concepts.3 When there was neither conceptual equivalency nor conceptual similarity between ICNP and CCC, the remaining ICNP concepts were manually clustered using 21 components (e.g., activity, medication, self-care) of the CCC framework. Any disagreements were further discussed with the ICN eHealth team and SabaCare until consensus was reached.

Results

Of 720 ICNP nursing diagnoses, the first step yielded 163 equivalencies (23%) between ICNP and CCC.6 In the second step, 328 ICNP diagnoses (46%) were identified as having conceptual similarity with CCC diagnoses – in all cases the meaning of the ICNP concept was narrower than that of the CCC statement. Lastly, the remaining 299 ICNP diagnoses were categorized into one or more of the 21 CCC care components. Table 1 presents example concepts selected from each step. In contrast, of 735 ICNP nursing interventions, 138 (19%) were exactly matched while the remaining 601 ICNP interventions were conceptually narrower or broader than CCC interventions.

Table 1. Example ICNP and CCC Concepts Selected from Each Step.
<table>
<thead>
<tr>
<th>Harmonization Step</th>
<th>ICNP Statement (2015 Release)</th>
<th>CCC Statement (Version 2.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEP 3. Concept clustering: ICNP concepts are grouped using the CCC component</td>
<td>10030192 Adherence To Medication Regime</td>
<td>G. Health Behavior</td>
</tr>
<tr>
<td></td>
<td>10030171 Adherence To Fluid Regime</td>
<td>F. Fluid Volume</td>
</tr>
</tbody>
</table>

**Discussion and Conclusion**

Our analysis of the two coding systems demonstrated structural differences between ICNP and CCC. ICNP is the largest compositional reference nursing terminology where each concept is defined formally using properties and associations with other concepts. This concept modeling process is necessary to prevent inconsistency and ambiguousness of an individual concept within the terminology system. An individual statement in CCC, however, represents a set of abstract concepts that collectively describe a nursing problem or action. For example, Medication Treatment in CCC is defined as ‘actions performed to administer/give drugs or remedies regardless of route,’ implying multiple activities (e.g., medication verification, medication administration and monitoring in ICNP) can be mapped to the CCC statement. While this structural and conceptual difference between ICNP and CCC was challenging to cross-map ICNP concepts and CCC statements, the 3-step harmonization process designed in this study has promoted to meet the goal of this project. Importantly, it became apparent that any attempt as universal one-to-one mappings between ICNP concepts and CCC statements will not reveal a complete solution for practical application. Rather the differences between the reference terminology (i.e., ICNP) and the classification system (i.e., CCC) encouraged the development of an alternative solution to mapping – the representation of ICNP concepts within the CCC conceptual framework. This harmonization work also demonstrated the versatility and flexibility of ICNP in complementing each system’s strength. Further study is needed to test the applicability of the 3-step process of harmonizing ICNP and other classification systems.

**Acknowledgements**

The authors would like to thank both ICN and SabaCare, Inc. for their support, and acknowledge the contribution of Dr. Virginia Saba to the work.

**References**

Comparison of Data Models used in Research Data Repositories for Electronic Phenotyping

Jeffrey G. Klann, PhD1,2,3; Vijay Raghavan, MS, MBA1; Michael Mendis, BS2; Douglas MacFadden, MS1; Sarah Weiler, PhD1; Kenneth Mandl, MD, MPH1,4; Shawn N. Murphy, MD, PhD1,2,3

1Harvard Medical School, Boston, MA; 2Partners Healthcare, Boston, MA; 3Massachusetts General Hospital, Boston, MA; 4Computational Health Informatics Program, Boston Children’s Hospital Boston, MA

Introduction

Research data repository systems and research networks are increasingly appearing, utilizing the vast quantity of information collected routinely by electronic health record systems. Each initiative requires data to be stored in a particular data model to support shared analytical tools. Several very different standard data models for clinical research are emerging. Each data model offers unique advantages and disadvantages in the storage, retrieval, and analysis of clinical data. Here we analyze three very popular but divergent approaches. Broadly, they differ most markedly in whether they focus on a single-table approach for storing data or a more traditional relational schema. The models include:

- **PCORnet Common Data Model (CDM).** The Patient Centered Outcomes Research Institute (PCORI)’s PCORnet is a rapidly developing collection of data research networks than presently span almost 80 clinical sites across the nation to perform large-scale comparative effectiveness research. [1] For pan-PCORnet queries, all networks support a common data model (CDM) developed by PCORnet’s Coordinating Center. This is derived from the Mini-Sentinel data model, which has increasing uptake in claims data analysis. [2]

- **Informatics for Integrating Biology and the Bedside (i2b2).** i2b2 is an open-source clinical data warehousing and analytics platform funded by the National Institutes of Health. [3] It is used at over 100 sites nationwide, including several PCORnet networks and the National Center for Advancing Translational Sciences (NCATS) Accrual to Clinical Trials (ACT) network. i2b2 uses a star-schema format, pioneered by General Mills in the 1970s and very popular in retail data warehouses. [4]

- **Observational Medical Outcomes Partnership (OMOP).** OMOP has been adopted by the Observational Health Data Sciences and Informatics (OHDSI) Consortium, a diverse, multi-stakeholder collaboration dedicated to providing robust analytical tools for research and quality improvement. [5] Because of its large community and an impressive suite of free analytical tools, the OMOP model is gaining traction in many research environments. Its data model is a hybrid of star schema and relational schema.

Here we focus on the critically important process of electronic phenotyping. [6] In this framework, we provide an analysis of each approach and one practical example: The Scalable Collaborative Infrastructure for a Learning Health System (SCILHS, pronounced “skills”). SCILHS is a growing network of health centers across the United States, presently covering over 10 million patients at eleven sites. SCILHS is a Clinical Data Research Network (CDRN) in PCORnet, as well as an i2b2-based network with a similar architecture to ACT. Therefore, it supports both CDM and i2b2.

Methods

CDM. CDM is designed with a traditional relational-database structure. The current release (v3) contains fifteen tables, each corresponding to a clinical domain (e.g., diagnoses, vitals, procedures, etc). The tables are wide, with many columns including both table keys (patient identifier, encounter identifier, etc) and additional details about each data element (e.g., primary diagnosis flag). New releases of the data model have occurred when new clinical elements are supported in analysis – for example, recent releases have added new domains (e.g., lab values) and made changes in data representation (e.g., smoking status).

i2b2. i2b2’s data model is designed to be highly adaptable and able to easily ingest data from various source systems without data transformation. As such, it has adopted a “star schema,” with one large “fact” table containing individual atomic observations. This is a narrow table with many rows per patient encounter. Local implementations develop concept hierarchies (called “ontologies”) that provide a window into the imported data. Consequently the data model is only modified when core features are added to the platform. Import of new types of data elements can be done directly into the fact table, and the ontology can be modified to make these data accessible to researchers. Analytical tools and data networks require conformance to a standard ontology, but this can be updated easily as network data availability changes, and it does not require changes to sites’ underlying data or schemas.
**OMOP.** OMOP is a hybrid approach that also provides a fact table (like i2b2) and domain tables (in the vein of CDM). Its schema is significantly more complicated than either of the other approaches (specifying 39 tables). Unlike CDM, the domain tables are in many cases derived values for specific analytical purposes (e.g., drug_era and visit_cost), and, like i2b2, raw data can be stored in the fact table. Also similar to i2b2, OMOP provides metadata tables providing information on terminology and concept relationships. Unlike i2b2, however, this terminology is standardized and not modifiable at each site. This requires data to be transformed to standard terminologies upon import, but demanding such standardization does enable a broad analytics suite.

**Results**

**Comparison of Approaches**

Each data model has significant strengths and weaknesses. PCORnet CDM is a ready-to-use analytical data model that is easy for researchers to understand and write analytical tools against. However, it is limited to only the data domains defined in the schema, and because clinical data is so diverse and source systems change so frequently, it is not a likely choice for robust electronic phenotyping algorithms. i2b2, on the other hand, is very flexible and has the proven capability to flexibly import data from a variety of source systems, even as source systems change. This may explain its popularity in many large research hospitals. i2b2’s ontology system allows dynamic query design and execution on local data, and, with mapping to standard ontologies, common analytical tools can be written. However, the star-schema is not very intuitive for researchers designing analytics, and analysts frequently prefer data transformed into a traditional column-oriented format like CDM instead. OMOP provides both tables for a standard analytical model and a fact table for dynamic support of new data elements. This hybrid approach is very promising. However, it does lose out on some of i2b2’s flexibility - due to requiring terminology mapping at time of import. This limits integration of new data types (e.g., mHealth data) to improve phenotyping algorithms.

**The SCILHS Network**

The SCILHS network is live at 11 sites across the United States, presently covering over 10 million patients. The SCILHS workflow uses i2b2 as the primary data warehouse, due to its flexible data ingestion and querying capabilities. We developed a robust mapping methodology to allow sites to map to a standard ontology without modifying the underlying data, which our sites have implemented. [7] The SCILHS network supports live preparatory-to-research queries using an ontology-driven graphical query tool. SCILHS also offers data in PCORnet CDM format for easy-to-understand data exports and compatibility with the larger PCORnet, using a transformation program based on the ontology mappings. This approach could be adapted to OMOP, so SCILHS could also take advantage of OMOP’s powerful analytical tools.

**Discussion**

Each analytical data model is optimized for some use-case, and transforming between them might be the best overall solution. [7] PCORnet CDM is a straightforward model that is very approachable to researchers and analytics developers. OMOP offers some data flexibility while providing a very powerful and expanding suite of analytical tools. For flexibly supporting complex, adaptive phenotyping algorithms, we find that i2b2 is superior – in that it supports diverse source data and is dynamically adaptable to new data element types and concepts, even when they are not represented in a standard terminology. The SCILHS network has chosen to use i2b2 for its primary data repositories, providing exports in other formats (presently PCORnet CDM).

**References**

Superiority of Carboplatin-Based Doublet Chemotherapy Regimens to Cisplatin-Based Doublet Chemotherapy Regimens among a Cohort of Medicare Patients Aged 70 and Older with Stage IV Non-Small Cell Lung Cancer

Elizabeth Lamont, MD, MS,1,2,3,4 Nancy L Keating, MD, MPH2,4,5 Christopher Azzoli, MD1,2 and Mary Beth Landrum, PhD5

1 Massachusetts General Hospital Cancer Center; 2 Department of Medicine, Harvard Medical School; 3 Department of Bioinformatics, Harvard Medical School; 4 Section of General Internal Medicine, Brigham and Women’s Hospital; and 5 Department of Health Care Policy, Harvard Medical School, Boston, MA

Background: During the period of study, platin-based chemotherapy doublets were established through phase III trials to be clearly superior to single agent chemotherapy regimens. For this reason, platin-based doublet chemotherapy regimens became the standard guideline recommended firstline therapy for patients with stage IV NSC lung cancer. Among the platin drugs, some evidence suggests that cisplatin-based chemotherapy doublet regimens are more effective than the less toxic carboplatin-based doublet regimens for patients with advanced NSCLC, but their effectiveness among elderly cancer patients who are treated in usual care settings is not known. We sought to compare cisplatin-based chemotherapy doublet regimens to carboplatin-based chemotherapy regimens among usual care elderly patients to determine if the findings from clinical trials were also evident in this understudied large subgroup of all patients who are diagnosed with stage IV NSC lung cancer in the United States.

Methods: We identified 13,406 elderly Medicare patients who were (1) diagnosed with stage IV NSCLC between the years 1995-2007 in one of the Surveillance, Epidemiology, and End Results (SEER) regions and (2) started treatment as an ambulatory patient with either a cisplatin-based or a carboplatin-based doublet chemotherapy regimen in the subsequent six months. Using propensity score weighting, we balanced the two treatment cohorts with respect to observable attributes. Table 1 shows our samples before and after propensity score weighting. We then estimated our key end points survival and post-treatment hospital-based health care use as a proxy for treatment-related toxicity.

Findings: Overall, patients treated with cisplatin-based doublets lived two weeks longer on average than patients treated with carboplatin-based doublets (i.e., 7.4 months vs. 7.0 months, p=0.05). For patients ≥70 years of age, first-line therapy with cisplatin-based doublet chemotherapy regimens was associated with increased post-treatment morbidity and no survival advantage compared to carboplatin-based doublet therapies.

Interpretations: Our results show that overall there was a clinically small survival advantage of marginal statistical significance for elderly Medicare patients with NSCLC who are treated with cisplatin-based doublet chemotherapy regimens compared to patients treated with carboplatin-based doublet chemotherapy regimens. However, the overall survival benefit appears to have been due to a larger survival advantage among the large subset of patients aged 66 through 69 years who lived approximately three and one half weeks longer than similarly aged patients treated with carboplatin-based doublet regimens (p < 0.01). There was no apparent cisplatin-associated survival advantage among patients 70 years and older. These findings are apparent in Figure 1. Among this latter group of patients, there were clinically and statistically meaningful differences in morbidity associated with cisplatin-based therapy. Patients who were treated with cisplatin-based doublet chemotherapy regimens were observed to use more (1) hospital-based health care and (2) more intensive hospital-based care following the beginning of chemotherapy than patients who were treated with carboplatin-based doublets.

These results advance clinical medicine in several ways. First and most importantly, we fill an existing void by providing “real world” estimates of morbidity and mortality outcomes in elderly Medicare patients with NSCLC according to cisplatin-based vs. carboplatin-based doublet therapy and according to tumor histology and patient age. These results may be immediately useful to oncologists and their older patients
as they discuss treatment options by providing more representative estimates of morbidity and mortality of two first-line treatment options. These real world estimates are of critical importance given that the average survival following both forms of doublet therapies (i.e., 7.0 months for carboplatin-based doublets to 7·4 months for cisplatin-based doublets) were approximately six weeks shorter than those reported in existing data like meta-analyses of clinical trials (i.e., 8·4 to 9·1 months respectively). Further, the small (< 2 weeks) and marginally significant survival advantage associated with cisplatin-based regimens in the larger cohort was clearly driven by the youngest patients, those 66-69 years of age who experienced a larger (> three week) survival advantage which was statistically significant. For those 70 years of age and older, there was no survival advantage, but clearly greater morbidity as represented by greater subsequent use and intensity of hospital-based health care. This suggests that carboplatin-based doublets are superior to cisplatin-based doublets for such patients through of minimization of treatment-related morbidity.

Figure 1. Forest Plot Comparing Relative Hazard of Death of Carboplatin- vs. Cisplatin-Based Doublets in Elderly Medicare Patients with Advanced NSCLC by Patient Age (N=13,406)

Legend: In this forest plot, unity (1.0) represents equality in patient survival between carboplatin-based vs. cisplatin-based doublet therapy among elderly Medicare patients with metastatic NSCLC. Deviation rightward represents superior survival associated with cisplatin-based doublet therapies and deviation leftward represents superior survival associated with carboplatin-based doublet therapies. NSCLC=non-small cell lung cancer
A Regional Cardiovascular Data Registry Linking Hospitals and Ambulances

James R. Langabeer II, PhD; Safa Fathiamini, MD MS; Tiffany Champagne-Langabeer, PhD; Jenny Mikhail, MPH; Wendy Segrest, MS; Raymond Fowler, MD, FACEP

1 The University of Texas Health Science Center, Houston, TX; 2 American Heart Association, Denver, CO; 3 The University of Texas Southwestern Medical Center, Dallas, TX

Introduction

Analyzing and reducing ‘total ischemic time’ is critical for improving the care and survival of acute myocardial infarction (MI, heart attack) patients as well as for clinical researchers focused on improving cardiovascular outcomes. Yet, an electronic health records system (EHR) maintains only part of the data, once patients arrive at their doors, and out-of-hospital data are generally unavailable. Physicians and hospitals must continually look for ways to reduce time delays in door-to-balloon (D2B) and system onset to arterial reperfusion (SOAR); and key to this is understanding the longitudinal record of care for cardiovascular patients prior to arrival. This study describes one unique longitudinal data registry developed for a regional myocardial infarction quality improvement project sponsored by the American Heart Association, specifically to share and link patient time and treatment data between 25 emergency medical services (EMS) agencies and 15 hospitals in Dallas Texas (AHA Caruth Initiative).

Objective

The objective of this study was to create an integrated, automated cardiovascular care data registry for one of the largest geographical regions in the country (Dallas Texas; the 9th largest city in the US). Dozens of commercial EHRs were in use by the participating 15 hospitals, and nearly 20 different custom patient care record (PCR) systems by the 25 EMS agencies, creating disparate data sources and formats that were impossible to assemble and analyze for clinical informatics purposes of measuring total ischemic time and identifying areas for improvement. Analyzing longitudinal patient-level data is difficult given that both sets of providers use different systems and codes to identify their patients, and in the absence of a master patient index (MPI), matching patient records across those systems is far from straightforward. Figure 1 shows the conceptual longitudinal linkage between the data sources.

Figure 1: Linking Ambulance and Hospital Records Reduces Treatment Times

Methods

De-identified data for MI patients who underwent primary percutaneous coronary intervention (PCI) in Dallas County, Texas were extracted in extensible markup language (XML) and shared securely by independent EMS agencies and hospitals monthly to the University of Texas Health Science Center researchers. A Microsoft SQL Server database was created with tables to store demographic, intervention, outcome, and time data. Automated scripts were developed to first scan the data for specific quality concerns (such as appropriate dates and values) and
then insert the data by source (hospital or agency) into the database. A probabilistic algorithm was written to help identify and match EMS to hospital records for each patient. Since the two data sets did not share a common patient identifier, the patient matching algorithm was based primarily on: a) approximate date and time of myocardial infarction; b) patient zip code; c) patient date of birth; d) hospital destination; and e) gender. After matching, an exception report alerted the system administrator of unmatched records for human review. Automated statistical analyses and reports were generated on a monthly and quarterly basis for all stakeholders in the region.

**Figure 2: Cardiovascular Registry Schema**

![Cardiovascular Registry Schema](image)

### Results

This cardiovascular care registry was developed to support analytical investigation of patient and provider level longitudinal time-based outcomes, which otherwise did not exist. We linked approximately 90% of all pre-hospital records against hospital records. After implementation of the registry, a new performance metric was created (termed Symptom Onset to Arterial Reperfusion), which allowed cardiologists, nurses, EMS agencies and others to manage key time-to-treat performance outcomes between ambulances and hospitals. The quarterly results were used for quality improvement processes. Teams were assembled to review findings, identify opportunities, and recommend changes to processes that might impact treatment times. Several key changes were made, such as allowing paramedics to remotely activate the catheterization lab from the ambulance. Challenges faced included financial, technology, data privacy, and organizational. The program resulted in a significant reduction in two key cardiovascular outcomes, door-to-balloon and total ischemic time, both between 10-15% improvements over a two year period. ²

### Discussion

Time is critical in emergency cardiovascular care. Registries which store longitudinal data on times and treatments allow quality improvement programs to identify collaborative strategies for reducing both pre- and in- hospital treatment delays. This study confirms that when EMS and hospital information is longitudinally shared, clinical informatics can produce useful analyses and help drive improvements in outcomes.

### References


An Automatic, Multi-Algorithm Approach to Classify Ventilator Waveforms

Monica Lieng¹, Brooks Kuhn, MD¹, Greg Rehm¹, Edward Guo¹, Jean-Pierre Delplanque, PhD¹, Nick Anderson, PhD¹, Jason Adams, MD, MS¹
¹University of California, Davis

Introduction

Mechanical ventilation (MV) is a life-saving intervention for critically ill patients. Approximately 30% of patients in Intensive Care Unit (ICU) environments receive MV¹. Due to variety in underlying disease and rapid changes in patients’ physiological states, events where breath delivery is inadequate for patient demand are common. These events, known as patient-ventilator asynchronies (PVA), have been correlated with structural injury to the lung², increased sedative use², prolonged MV duration³,⁴ and, in one observational study, increased mortality⁵.

Quantitative analysis of PVA for research and improvement of patient care has historically been hindered by low data sampling intervals and manual paper-based recording methods. Several previous studies have demonstrated the feasibility of algorithmically detecting PVAs using a computerized system⁵,⁶. Most of these studies have looked at only one sub-type of PVA, with only one study reporting on the detection of more than two subtypes.⁵ Furthermore, these studies have been limited to brief sampling intervals, small sample sizes, use of un-validated proprietary software, and lack of the ability to filter out common clinical artifacts.

Misclassification of PVA events may occur due to morphologic similarities between common clinical artifacts (e.g. cough, endotracheal tube suctioning) and individual PVA subtypes. For example, patients receiving endotracheal tube suctioning for mucous removal may generate waveforms that share morphologic features with double triggering asynchrony. In a single-event algorithmic approach, such artifacts may be mistaken for clinically significant asynchronies which may in turn over-represent the prevalence of true PVA events or result in alert fatigue if applied in a clinical decision support application. To address these issues, we describe an integrated multi-event algorithm approach for simultaneously detecting a variety of PVA subtypes in clinically-derived ventilator waveform data based on a heuristic evaluation framework derived from the clinical rules used for the detection of PVA by bedside clinicians.

Methods

Continuous flow and pressure waveforms were acquired from both healthy volunteers and ICU patients on Puritan Bennett 840 mechanical ventilators using an unobtrusive wireless peripheral device. Classification algorithm development used a combination of simulation-based and clinically-derived waveforms. Waveforms were visualized using a novel visualization platform powered by Flask and Dygraphs and algorithm performance was compared to gold standard reference data sets produced from two-physician consensus annotations. All algorithms were implemented in Python 2.7. Clinical metadata were extracted from raw waveform data to facilitate PVA detection. These included common calculations such as expiratory and inspiratory tidal volume, respiratory rate and peak inspiratory pressure. Similar to previous work⁵,⁶, we estimated tidal volume by integrating the area under the inspiratory and expiratory flow-time curves. Expert physicians defined the heuristic rules by which common PVA subtypes and clinical artifacts are classified. These rules were translated into pseudocode and then subtype-specific algorithms in Python. Additional heuristic algorithms were created to parse the “raw” output of the individual classification algorithms in order to distinguish between PVA subtypes and common artifacts. For example, if a given breath is classified as both a suction event and a double trigger asynchrony, our integrated multi-algorithm approach favors the suction event and removes the detection of the double trigger to prevent the misclassification of a PVA-like artifact as a true PVA event. We then compared the sensitivity, specificity and positive predictive value (PPV) of our individual PVA subtype classification algorithms to that of our integrated multi-event algorithms.

Results

Data were collected from 17 patients and clinicians sampled 15 minute regions with high PVA and artifact density. The pooled data contained 5075 breaths, where 1179 were pathologic (double trigger or breath stacking) and 182 were artifacts (multiple trigger, cough, ventilator disconnect or suction). Overall, algorithm performance appears to have improved for both double trigger and breath stacking (Table 1). Notably, PPV for the double trigger and breath stacking improved from 66.7% to 91.1% and 92.4.7% to 95.2%, respectively. For example, patient 7 had improvements in accuracy, specificity and PPV for double trigger asynchronies. In the single-event algorithm, 33 false-positive double triggers were detected where only 28 true positives existed, representing more than a 2-fold overestimation of the prevalence of double trigger events when using the single-event classification algorithm. After implementing the multi-event algorithm, the number of false positive double trigger events reduced to 2.
**Discussion**

Our results suggest that implementing a multiple-event detection algorithm improves PVA sub-algorithm specificity without substantially compromising sensitivity. In particular, the double trigger sub-algorithm saw improved performance with the integration of clinical artifact filtering sub-algorithms. While preliminary, our findings highlight the need for clinical artifact detection and filtering approaches for the analysis of streaming waveform data. We have shown that single-event detection can result in a substantial over-estimation of the true prevalence of clinically meaningful events when artifacts are not detected and removed from analysis. In addition to the effects that artifact detection and removal may have on the study of PVA in the ICU, multi-event detection algorithms stand to improve the efficacy of clinical decision support (CDS) analytics by reducing alert fatigue and potentially improving the prognostic performance of CDS systems to identify those patients truly at risk of adverse outcomes. Furthermore, reducing the false positive rate could decrease potentially unnecessary and harmful changes in ventilator settings that may lead to additional patient distress.

Our study is limited by the relatively small sample size of our derivation cohort and relatively low PVA and artifact event rates despite manually enriching for high event rate regions of interest (ROI). While our manual ROI selection process may have introduced bias into algorithm performance, manual ROI selection was felt necessary given large numbers of normal breaths and the absence of validated available waveform analysis software with which to select ROIs. For proof of concept, we focused on a limited subset of PVA and artifact types and it is possible that multi-event algorithm performance may change when additional PVA and artifact sub-algorithms are added in the future.

This approach demonstrates the ability of a multi-algorithm framework to analyze large volume streaming waveform data acquired from mechanically ventilated ICU patients. Future work will focus on analysis of multi-event algorithm performance in a separate validation cohort, use of our analytics to study the relationship between PVA and clinical outcomes, and development of a CDS system using our multi-event algorithm to deliver actionable near real time information to critical care care providers to improve the management of patients with acute respiratory failure.

**References**

Technology, Incentives, or Both?
Factors Related to Level of Hospital Health Information Exchange

Sunny C. Lin, M.S. 1, Jordan Everson, M.P.P. 1, and Julia Adler-Milstein, Ph.D. 1,2
1School of Public Health, 2School of Information
University of Michigan, Ann Arbor, MI

Introduction
A key component of Stage 2 meaningful use (MU) is using EHRs to share information to improve care coordination. Hospitals must generate a summary of care record (SCR) and transmit it for at least 50% of care transitions, with 10% transmitted electronically. The electronic transmission threshold was set low because of concern that hospitals may not have the capabilities to engage in high levels of HIE. However, hospitals with HIE capabilities may still limit exchange volume given concerns that HIE may put them at a competitive disadvantage. We therefore sought to determine the extent to which the level of electronic SCR transmission was related to hospital capabilities, incentives, or both.

Methods
We combine Stage 2 MU attestation data with AHA Annual and IT Supplement survey data. Capability factors included whether the hospital had a basic or comprehensive EHR, potential trading partners (percent of hospitals in the market that attested to Stage 2 MU), and active participation in a Health Information Organization. We chose incentive factors based on theories from resource dependency and institutional theory, which suggest that hospitals may share more information if they are dependent on partners in the communities (e.g. participation in capitated or shared risk programs, or members of a hospital system or network), but may share less information if they are struggling to maintain demand for their services or have an organizational emphasis on revenue generation (e.g. hospitals in more competitive markets, hospitals with less market share, or for-profit hospitals). (1) Hospitals may also see a benefit in exchanging more information to maintain legitimacy if their mission is to serve the community (e.g. non-profit hospitals). (2) We used multivariate OLS regression with state fixed effects to analyze the relationship between capability and incentive measures and the percent of care transitions with an SCR transmitted electronically. To assess whether factors were uniquely related to the level of electronic transmission, we repeated the analysis using percent of transitions where an SCR was sent in any mode and percent of SCRs sent electronically. Controls include: hospital size, teaching status, percent Medicare and Medicaid days, geographic location, hospital beds in county, and hospital-ambulatory integration. Hospitals included in the analyses were all non-federal, general acute-care hospitals in the United States that attested to Stage 2 Meaningful Use as of April 2016 and responded to 2014 AHA IT Supplement survey (N=1778).

Results
Full regression results are presented in Table 1. For-profit ownership, compared to government, was associated with a 9.570 decrease in percent of discharges with an electronic SCR (p<0.001). Market share and competition were not significant. System membership was associated with a 4.123 increase in percent of discharges with an electronic SCR (p<0.001). However, network membership did not have a statistically significant relationship with percent of discharges with an electronic SCR. Compared to less than basic EHR, having a comprehensive or basic EHR were associated with a 4.370 increase in percent of discharges with an electronic SCR (p=0.002). The other two capability variables were not significant. Of these statistically significant factors, only for-profit ownership was also associated with any-mode SCR transmission; however, for profit ownership was associated with a 12.008 increase in the percent of discharges with any SCR (p<0.001). This suggests that for-profit ownership is associated with higher SCR transmissions overall, but lower electronic SCR transmissions specifically.

Discussion
This is one of the first national studies to examine a precise measure of the level of HIE, instead of simply HIE capabilities. We sought to understand what factors influence the level of exchange after the capability to exchange data is already determined. We found that the percent of care transitions with an electronically transmitted SCR was related to two incentive-related factors (ownership and system membership) and one capability measure (having a more advanced EHR). While our hypothesis suggests that this is because hospitals with emphasis on revenue generation and those with greater dependence on community partners may be incentivized to share less or more information, respectively, further research into the mechanisms behind this finding is warranted. Our study suggests that capability factors as well as incentive factors may
be influencing the extent to which hospitals are engaging in HIE. Achieving high thresholds of exchange in the future will therefore require addressing both domains.

Table 1. Relationship between Hospital Characteristics and Percent of Care Transitions where an SCR was Transmitted.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Primary Analysis</th>
<th>Secondary Analysis¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outcome 1: Percent Discharges with Electronic SCR</td>
<td>Outcome 2: Percent Discharges with Any SCR</td>
</tr>
<tr>
<td>Constant</td>
<td>34.091*** (5.465)</td>
<td>83.636*** (4.40)</td>
</tr>
<tr>
<td><strong>Incentive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HHI</td>
<td>0.797 (6.667)</td>
<td>-8.096 (5.031)</td>
</tr>
<tr>
<td>Market Share</td>
<td>0.074 (0.086)</td>
<td>0.060 (0.065)</td>
</tr>
<tr>
<td><strong>Ownership (Ref: Gov’t)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non Profit</td>
<td>-2.404 (1.642)</td>
<td>0.317 (1.357)</td>
</tr>
<tr>
<td>For Profit</td>
<td>-9.570*** (2.312)</td>
<td>12.008*** (1.877)</td>
</tr>
<tr>
<td>% Rev Capitated</td>
<td>0.005 (0.219)</td>
<td>0.049 (0.162)</td>
</tr>
<tr>
<td>% Rev Shared Risk</td>
<td>-0.016 (0.096)</td>
<td>-0.106 (0.072)</td>
</tr>
<tr>
<td>Network Membership</td>
<td>-1.632 (1.176)</td>
<td>-0.747 (0.938)</td>
</tr>
<tr>
<td>System Membership</td>
<td>4.123** (1.344)</td>
<td>-0.647 (1.096)</td>
</tr>
<tr>
<td><strong>Capability</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EHR Capability (Ref: &lt; Basic)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic EHR</td>
<td>1.444 (1.510)</td>
<td>0.475 (1.30)</td>
</tr>
<tr>
<td>Comprehensive EHR</td>
<td>4.370** (1.573)</td>
<td>-0.051 (1.345)</td>
</tr>
<tr>
<td>% Hospitals that Electronically Exchange with other Hospitals</td>
<td>0.047 (0.030)</td>
<td>-0.027 (0.024)</td>
</tr>
<tr>
<td>Active HIO Participation</td>
<td>1.779 (1.308)</td>
<td>2.866** (1.056)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size (Ref: Large)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>-1.478 (1.645)</td>
<td>-3.819** (1.314)</td>
</tr>
<tr>
<td>Medium</td>
<td>-2.770 (2.698)</td>
<td>-4.162* (2.070)</td>
</tr>
<tr>
<td>Teaching</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accredited</td>
<td>-1.650 (1.485)</td>
<td>-1.545 (1.148)</td>
</tr>
<tr>
<td>COTH Member</td>
<td>-8.060** (2.732)</td>
<td>-4.298* (2.115)</td>
</tr>
<tr>
<td>% Rev Medicare</td>
<td>-0.016 (0.043)</td>
<td>-0.023 (0.035)</td>
</tr>
<tr>
<td>% Rev Medicaid</td>
<td>0.040 (0.053)</td>
<td>-0.048 (0.044)</td>
</tr>
<tr>
<td>Urban Code (Ref: Metro)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Micro</td>
<td>-0.731 (1.756)</td>
<td>-0.585 (1.411)</td>
</tr>
<tr>
<td>Rural</td>
<td>3.519 (2.021)</td>
<td>3.604* (1.615)</td>
</tr>
<tr>
<td>Population in HSA (in thousands)</td>
<td>0.000211 (0.001)</td>
<td>0.0 (0.001)</td>
</tr>
<tr>
<td>Hospital Beds in HSA</td>
<td>0.513 (0.553)</td>
<td>0.498 (0.379)</td>
</tr>
<tr>
<td>Physician-Hospital Integration</td>
<td>0.110 (0.413)</td>
<td>-0.282 (0.335)</td>
</tr>
<tr>
<td><strong>State Fixed Effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Included</td>
<td>0.040</td>
<td>0.1173</td>
</tr>
<tr>
<td>Included</td>
<td>0.2915</td>
<td>0.0610</td>
</tr>
<tr>
<td>Included</td>
<td>0.0633</td>
<td>0.1270</td>
</tr>
<tr>
<td>n</td>
<td>1778</td>
<td>1030</td>
</tr>
</tbody>
</table>

Standard errors in parentheses, legend: *p <0.05, **p <0.01, *** p< 0.001

¹Data for the secondary analysis comes from the unmodified Stage 2 meaningful use program

References


86
Designing a Tablet Tool for Complex Patients to Prioritize Discussion Topics during In-Person Primary Care Visits
Courtney Lyles, PhD, Neetu Chawla, PhD, Christine Kowalski, MPH, Deanna McQuillan, MA, Andrea Altschuler, PhD, Elizabeth Bayliss, MD, MSPH, Michele Heisler, MD, MPH, & Richard Grant, MD, MPH

Introduction: Complex patients (i.e., those with several concurrent chronic diseases) experience higher healthcare costs and poorer health outcomes, partly driven by limited amounts of time to discuss all of their health issues with their doctors. User-friendly technology may improve communication between complex patients and their primary care providers by helping prioritize health concerns during brief in-person visits. We employed user-centered design using in-depth qualitative methods to create a tablet tool to assist complex patients with setting discussion topic priorities at upcoming visits – as this population that is comprised of mostly older adults is not often the focus of technology design. We outline here the main findings from this qualitative work, as well as the specific design implications that emerged.

Methods: We conducted semi-structured interviews with 40 patients and their 17 primary care providers across three sites: Kaiser Permanente Northern California (n=12 patients and 5 providers), Kaiser Permanente Colorado (n=13 patients and 8 providers), and University of Michigan, Ann Arbor (n=15 patients and 4 providers). All patients had 2 or more chronic conditions. In-depth interviews across sites lasted approximately 45 minutes. We then conducted three 60-minute iterative focus group sessions (all at KPNC using the same recruitment criteria, n=12 with repeat attendance by most patients) to iterate a paper-based prototype of the tablet tool with screenshots of the application and scripts of the audio clips. All interview and focus groups discussions were audiorecorded and transcribed for analysis. For the interview analysis, three coders (one at each site) collectively determined the primary discussion topics, and then each of these coders pulled all relevant transcript excerpts related to those topics. For the focus group analysis, we reviewed all design/content modification recommendations that emerged from the group design sessions. We organized recommendations into themes and sub-topics and linked these exemplary quotes to the major changes during our design process.

Results: The mean age of participants in the baseline interviews was 65-73 years old across the sites; 58% of participants were women. Focus group participants were primarily women (71%), with an average age of 71. Key themes across all qualitative work included: 1) streamlining the tablet tool design (including for those with lower technology proficiency), 2) improving tool content, and 3) recommending appropriate patient subgroups for the final tool (specific findings are in Tables 1-3). A total of 25% of patients in the baseline interviews explicitly stated that they were not interested in using this tool during their upcoming visits, for technology proficiency or lack of fit to their ongoing care processes (e.g., already made lists for providers).

<table>
<thead>
<tr>
<th>Table 1. Patient Interview Findings</th>
<th>Sub-topic</th>
<th>Example quote(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Challenges for patients with limited technology proficiency</td>
<td>Patient lacks skills to use a tablet</td>
<td>“I finally bought an iPad not too long ago and I still don’t even know how to use that. I don’t know much about computers.”</td>
</tr>
<tr>
<td></td>
<td>Suggestions to make tablet easier to use</td>
<td>“Can it be like something like, on my phone, you have ‘speak’ and ‘talk’?”</td>
</tr>
<tr>
<td>2. Patient recommendations for content of Pre-Visit Tool</td>
<td>Elicit simple lists of ongoing or new problems</td>
<td>“What am I here for?” You know, “What is my major complaint?”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I don’t mind checking when you give me a pad and it says ‘Check off what’s wrong with you.’”</td>
</tr>
<tr>
<td></td>
<td>Provide a way to bring up topics patients might overlook</td>
<td>“Maybe some things that I wouldn’t think of that would be related to what I’m dealing with.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I may need, like, you know, a trigger: Oh, yeah, you know, I did want to ask about this.”</td>
</tr>
</tbody>
</table>
| | Provide opportunities for | “It depends on how the information is presented to a
These qualitative findings informed the following major design decisions: 1) making the tablet tool accessible for patients who lacked technical proficiency by reducing the quantity and complexity of text presentation (e.g., adding 3 <30-second audio clips explaining/coaching how to use the tool and removing all scrolling and extra text from each screen), and 2) ensuring specific content choices based on patients’ and providers’ personal experiences (such as the ability to raise private or sensitive concerns), and 3) targeting new patients for using the tool in future work since they were seen to have the most communication needs. Our final tool had text at the 6th grade reading level, used 6 screens total with sufficient space for free text entry.

**Conclusions:** Applying user-centered design methods, we substantially modified our initial design concept based on robust input from patients and physicians. The final tool was simpler to use while also providing richer choices and more engaging layouts. Rigorously documenting this adaptation process is a critical step of health informatics research in order to generalize knowledge about how the user-centered design process works in real-world scenarios, especially for older adults with multiple chronic conditions who are often left out of technology design processes. In future work, we see the potential to use even more patient data from the EHR and other sources to personalize the information within the tablet tool even further.
POE: A Pathology Extraction Tool for Finding Attribute-Value Pairs in Glioma Pathology Reports.

Veronica E. Lynn, BA¹, Niranjan Balasubramanian, PhD¹, Tahsin Kurc, PhD¹, Joel Saltz, MD, PhD¹, Rebecca Jacobson, MD, MS²,
¹Stony Brook University, Stony Brook, NY; ²University of Pittsburgh, Pittsburgh, PA

Introduction
Histopathologic features, identified by direct examination of neoplasms, play a critical role in cancer diagnosis. Diagnostic classifications, such as the WHO classification for glial neoplasms, provide guidance on the sub-classification and grading of cancer. For example, a high degree of cellularity and necrosis is associated with higher grade and worse outcomes in glial tumors. New integrative approaches that combine information from histologic, imaging and genomic features have the potential to advance methodologies for cancer classification[1] and will likely necessitate development of new classification schemes. These approaches will require scalable extraction of semantically rich features, potentially in very large datasets.

At present descriptions of histopathologic features are available in free text surgical pathology reports, and natural language processing is needed to extract them. In this case study, we develop and evaluate a pathology extraction tool, POE, which builds on the TIES framework[2], for Glioma pathology reports. Many existing NLP tools [2,3,4] identify medically relevant entities drawn from medical terminologies; TIES in particular provides a carefully hand-crafted set of cancer-specific entities. Tools such as TIES provide additional markup including negative observations and findings which have diagnostic value. In addition to these valuable annotations, we also need entity-attribute relationships to capture salient morphologic features, and the grading of observations, which has critical diagnostic value. For example, we may want to know whether cellularity was high or low to assess whether the specimen is a low or high-grade glioma. While it is relatively easy to find modifiers, determining which entity they modify is difficult. The terse nature of the reports means there are often many possible candidates, making this a hard task. To address these challenges, POE uses i) a feature classifier to identify relevant histopathologic features, and ii) a feature grade extractor that uses a syntactic dependency parser to locate the modifiers for each feature mention and map it to a grade value.

Method
On its own, TIES processes documents to identify entities and annotates their text spans with their concept name, concept code, and negation information. POE extends this by using the entity annotations to identify specific morphological features and link them with modifiers, which are then mapped to a grade value. This process is explained in more detail in Figure 1. In total, POE can extract 12 features, chosen due to being commonly considered when making a diagnosis, which can be assigned values based on severity (absent, present, mild, mild to moderate, moderate, moderate to high, or high) or shape (round, round to oval, or oval). Some examples are given in Figure 1.

Results
We evaluated POE using 473 pathology reports (250 glioblastoma, 223 low-grade glioma) collected from the Cancer Genome Atlas (TCGA). For these reports, the final diagnosis is given but the feature values must be extracted from the text. We therefore manually annotated 213 of these reports (132 glioblastoma, 81 low-grade glioma) to obtain the feature values.
POE’s feature identification classifier was evaluated through cross validation on the manually annotated set. It was able to determine if a feature was mentioned within a report with an F1 of 87.2. We also evaluated POE’s extractor for its ability to assign values for each feature. Across all possible values, POE obtained an accuracy of 52.2% at assigning the correct grade, a nine percent improvement (statistically significant) over a nearest-modifier heuristic. The drop in accuracy reflects the difficulty of assigning a grade compared with identifying feature mentions. Grade assignment typically requires a deeper semantic understanding of the text, necessitating a more sophisticated approach.

We demonstrate the utility of POE’s extracted features by using them to train a random-forest classifier to predict glioma type. The classifier is evaluated through cross validation using the full 473 reports, all of which have been automatically annotated using POE. We achieve an accuracy of 76.3% (80.4% glioblastoma, 71.7% low-grade glioma), compared to a random chance baseline of 52.9%.

Discussion

Our preliminary evaluations show that POE’s extensions to TIES can effectively extract relevant histopathologic features from glioma pathology reports. The evaluations also show that the extracted histopathologic features have some diagnostic value, motivating investigations into their utility for assessing new classification schemes. Effective extraction of histopathologic features from free-text reports is a precursor to determining how these features relate to each other, to image-based features and with outcomes. It will be a critical component in developing robust, integrative approaches for evaluating new classification schemes and designing effective therapies. Our next steps include (i) addressing extraction issues which stem from stylistic variations and the terse nature of the reports, and (ii) scaling the feature identification for other types of cancers. To promote further investigations in the community, we have released a dockerized version of POE (https://hub.docker.com/r/sbubmi/pathomics_morphex) that can produce a list of histopathologic feature-value pairs from pathology reports.

References

Development of an Oncology Subset of SNOMED CT Based on Patient Notes

Sina Madani, MD, PhD1; Jerry Henderson, MD1; Kin Wah Fung, MD, MS, MA2
1The University of Texas MD Anderson Cancer Center, Houston, TX
2National Library of Medicine, Bethesda, MD

Abstract

MD Anderson Cancer Center (MDA) is one of the world’s largest institutions involved exclusively in cancer care, research, and prevention. More than 120,000 clinical transcribed documents are added to the MDA EMR system on a monthly basis. We used natural language processing methods to generate a subset of SNOMED CT concepts that are frequently documented as cancer diagnoses in patient notes.

Introduction

SNOMED CT is gaining momentum as an international clinical terminology as the membership of the International Health Terminology Standards Development Organization (IHTSDO) tripled from its inception in 2007 to 28 countries. In the U.S., SNOMED CT is the designated terminology for the problem list and procedures according to the Meaningful Use of the electronic health record incentive program, as well as for transmission of data to cancer registries. Similar to the Clinical Observations Recording and Encoding (CORE) Problem List Subset of SNOMED CT, a list of commonly used SNOMED CT concepts in oncology will save effort in mapping local terms to SNOMED CT, reduce variability in data capture and enhance data interoperability. We describe our experience in the creation of an oncology subset based on natural language processing (NLP) of patient data in a large cancer treatment center.

Methods

We retrospectively analyzed the content of more than 15 million clinical narratives entered in MD Anderson Cancer Center legacy EMR system using MetaMap NLP framework to extract active clinical problem and cancer-related concepts at the patient level. After careful examinations of various note types and consulting with subject matter experts, we decided to target the last ten instances of the targeted clinical notes for all patients inside our EMR repositories. The targeted note types included Discharge Summary, Emergency, History & Physical, Consultation, Primary Medical Evaluation, Progress, and Clinical notes. We defined a “Problem” or “Cancer Disease” as something that required a plan for “diagnosis” and/or “management”. Therefore, from the selected note types mentioned above, all section headers related to the Assessment & Plan and Diagnosis sections, including subsections such as “Cancer Diagnosis”, together with their contents were extracted and analyzed by MetaMap v2014. We empirically set the threshold of MetaMap output to a score of 580. We restricted the UMLS targets to the “Disorders” semantic group, which included the semantic type “Neoplastic Process” for cancer-related concepts. We have also developed a post-processing module that prevented particular trigger strings within the narratives being mapped incorrectly to UMLS concepts, specifically, when the target concept was represented in an abbreviated format (like “gist” for Gastrointestinal Stromal Tumor). Such functionality has since been added to the latest version of MetaMap (v2016). One Extensible Markup Language (XML) file was generated for each patient note and serialized into a relational database. Phrases that MetaMap failed to map were stored separately for review. We applied the CORE Problem List subset (v2015) as an initial filter for removing unwanted and/or irrelevant concepts from the MetaMap output. Two trained physicians reviewed all concepts that were not represented in the SNOMED CORE subset and identified the ones that were relevant to cancer diagnosis. We calculated concept usage index by the occurrence of a cancer concept in a patient’s problem list by the total number of recorded cancer across all patients. To evaluate our NLP pipeline, we manually reviewed the output in two random patient samples (one for
general problems and the other for cancer diagnoses) to calculate the standard NLP performance metrics (precision, recall, F-measure).

**Results**

Based on the selection criteria mentioned in the method section, 554,801 unique patient records associated with 2,998,322 notes and 3,404,575 section headers were processed in the MetaMap pipeline. We identified 563 synonyms for the two categories of the target section headers. The performance metrics of the section header identification algorithm calculated as 97%, 99%, and 98% for precision, recall, and F-measure respectively. More than 3.7 million instances of cancer concepts corresponding to 2,698 unique concepts were extracted (Table 1).

<table>
<thead>
<tr>
<th>UMLS CUI</th>
<th>Concept</th>
<th>SNOMED Code</th>
<th>Usage Index by</th>
<th>by patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0006142</td>
<td>Malignant tumor of breast</td>
<td>254837009</td>
<td>6.7</td>
<td>6.4</td>
</tr>
<tr>
<td>C0025202</td>
<td>Malignant melanoma</td>
<td>372244006</td>
<td>4.9</td>
<td>4.3</td>
</tr>
<tr>
<td>C0024299</td>
<td>Malignant lymphoma</td>
<td>118600007</td>
<td>2.7</td>
<td>2.4</td>
</tr>
<tr>
<td>C0376358</td>
<td>Malignant tumor of prostate</td>
<td>399068003</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>C0007131</td>
<td>Non-small cell lung cancer</td>
<td>254637007</td>
<td>2.1</td>
<td>2.2</td>
</tr>
</tbody>
</table>

**Table 1.** Top 5 cancer concepts and their usage index by instance, note, and patient frequencies.

NLP performance evaluation for the general problem list concepts (Disorders semantic group) on twenty randomly selected patients showed 94%, 90%, and 92% for recall, precision, and F-Measure respectively. We also evaluated performance metrics for only Neoplastic Process semantic type on a separate group of twenty randomly selected patients and calculated 100%, 83%, and 90% (recall, precision, and F-Measure). Nine cancer diagnosis discovered by MetaMap in this group (like Carcinosarcoma vs. Ovarian Carcinosarcoma) were considered as “too general” by the evaluating subject matter experts.

**Discussion**

We showed that it is feasible to use MetaMap to extract cancer-related SNOMED CT concepts from narrative patient notes. The oncology subset will be made available for download through NLM’s website by any user with a SNOMED CT license. While the subset is not meant to be exhaustive, it can be used as a starter set as it is expected to cover the majority of SNOMED CT concept needed in most cancer treatment institutions. During the creation of the subset, we encountered some concepts that had considerable usage but were not in SNOMED CT. We would review the unmapped concepts for validity and submit them to IHTSDO as suggested additions to SNOMED CT. Furthermore, we plan to use NLP outputs for a gap analysis between a vendor’s terminology and the internally identified cancer synonyms from MDA corpora for the enrichment of the existing provider friendly terminology incorporated within Epic system.

**References**

3. Fung KW, Xu J. An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT. J Am Med Inform Assoc. 2015;22(3):649-58. Epub 2015/03/01.
Protein Biomarker Discovery and Ranking for Early Detection of Necrotizing Enterocolitis in Neonates

Subramani Mani MBBS PhD\textsuperscript{1}, Daniel Cannon MS\textsuperscript{1}, Karri Ballard PhD\textsuperscript{2}, Robin Ohls MD\textsuperscript{1}

\textsuperscript{1}University of New Mexico Health Sciences Center, Albuquerque, NM, 87131
\textsuperscript{2}Myriad Rules Based Medicine, Austin, TX, 78759

Necrotizing enterocolitis (NEC) is a serious inflammatory condition of newborns and infants with high mortality and morbidity and affects premature infants disproportionately. Small circulating blood volume in premature and low birth infants at higher risk of NEC and the need to minimize blood draw calls for the development of novel approaches for early detection of NEC. Developing informatics methods for protein biomarker discovery and ranking to facilitate early detection and intervention in NEC is a definitive clinical need as currently there is no specific diagnostic test that is available to clinicians in a neonatal intensive care setting. Predictive biomarkers are needed for early detection and to formulate treatment strategies for improved outcomes. Using a sample of 118 infants (24 cases and 94 controls) we identified the top 15 predictive protein biomarkers for NEC and ranked them.

Introduction and background: The recent publication titled “Improving Diagnosis in Healthcare” from the National Academies of Sciences, Engineering, and Medicine drives home the message that the delivery of healthcare has a significant blind spot, that is, diagnostic errors—inaccurate or delayed diagnoses [1]. Unfortunately this holds particularly true in the domain of necrotizing enterocolitis of neonates.

Necrotizing enterocolitis is a serious problem in neonatal intensive care units, particularly in extremely low birth weight (ELBW) infants. It affects 10\% of infants born <29 weeks gestational age [2, 3]. Nearly one half of NEC cases die, and many develop sepsis [4, 5]. Immaturity is clearly a causal factor [2], but clinical factors alone do not predict risk of NEC. About half of NEC cases occur 7 to 21 days of life, but timing of onset varies across neonatal intensive care units [2]. Predictive biomarkers are needed for early detection and to guide treatment strategies. Despite the frequency with which life-threatening NEC is encountered in the new born intensive care unit (NBICU), a rapid, sensitive, and specific diagnostic test remains elusive. Blood cultures are frequently negative, despite clinical symptoms that reflect significant inflammatory response, circulatory collapse, neutropenia, and thrombocytopenia. The identification of novel biomarkers that can provide a quantitative risk of development of NEC before the onset of symptoms will facilitate earlier intervention and improved outcomes [2].

Study objective: Our primary objective was to build predictive models and identify a set of ranked protein biomarkers for rapidly and reliably detecting necrotizing enterocolitis in infants.

Methods: We enrolled 118 eligible VLBW infants (gestational age \leq 32 weeks, birth-weight \leq 1500 grams, postnatal age \geq 120 hours) over a five year period from 2007 to 2012. Based on current best practice at NBICU, we assigned gold-standard diagnostic labels to the infants retrospectively using all available information. We categorized 24 infants as cases (NEC) as they satisfied Bell’s criteria for stage II and above [6] and the remaining 94 as controls. The Bell’s staging system for NEC is based on historical, clinical and radiographic criteria with stage I labelled as suspect and stages II and III considered definite and advanced NEC respectively. We used serum samples collected from each infant over a 21-day period to perform a focused proteomic assay of 90 potential biomarkers suspected to play a role in infection and inflammation. The quantitative proteomic assay was performed by Myriad RBM using a customized implementation of the Lumines xMAP technology, a microsphere-based multiplexed immunoassay platform. Using stratified 10-fold cross-validation, we evaluated a variety of machine learning methodologies including Random Forest, C4.5, ADT, logistic regression, and Naïve Bayes. Using Random Forest variable importance modeling we also ranked the predictive proteomic biomarkers. First we did this context free using only the proteins as potential predictors for modeling. Subsequently we also included the available clinical and lab parameters such as blood glucose, temperature, pulse rate, respiratory rate and repeated our predictive modeling experiments. The study protocol was approved by IRB and written, informed parental consent was obtained.

Results: Random Forest and logistic regression were the best two performing algorithms. Random Forest had the overall best performance with an Area under the curve (AUC) of 0.826, sensitivity of 0.817, specificity of 0.676, positive predictive value (PPV) of 0.391 and negative predictive value (NPV) of 0.935. See Table 1. Random Forest variable importance ranking identified 15 informative biomarkers which are shown in Table 2. Addition of the clinical parameters improved the predictive performance of Random Forest resulting in an AUC of 0.845.

Discussion and Conclusion: We have proposed an informatics approach based on machine learning to identify and rank the most predictive protein biomarkers for early detection and possible intervention in Necrotizing enterocolitis. The specific biomarkers TNFR2, IL-6 and IL-10 are known to be associated with inflammation and
infection. However, these are not routinely used in clinical practice. The role of other predictive biomarkers we identified have not been clearly established in NEC based on published studies. Endothelin-1 is a potent vasoconstrictor peptide which has been implicated in cardiovascular disease but plausibly can play a role in the pathogenesis of NEC due to its vasoconstrictive and pro-inflammatory actions. While additional validation is necessary before contemplating clinical use, our results suggest the potential utility of a proteomics-based diagnostic tool for early detection of NEC in preterm infants.

Acknowledgments: We thank the Thrasher foundation and the CTSC for funding support for the study.

Table 1: Predictive performance of Logistic regression and Random forest algorithms for NEC using 10 x 10-fold cross validation. PPV: Positive predictive value, NPV: Negative predictive value, s.d: standard deviation.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>AUC (s.d)</th>
<th>Sensitivity (s.d)</th>
<th>Specificity (s.d)</th>
<th>PPV (s.d)</th>
<th>NPV (s.d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression (proteins only)</td>
<td>0.820 (0.025)</td>
<td>0.621 (0.039)</td>
<td>0.847 (0.023)</td>
<td>0.511 (0.046)</td>
<td>0.897 (0.011)</td>
</tr>
<tr>
<td>Random forest (proteins only)</td>
<td>0.826 (0.007)</td>
<td>0.817 (0.020)</td>
<td>0.676 (0.013)</td>
<td>0.391 (0.013)</td>
<td>0.935 (0.007)</td>
</tr>
<tr>
<td>Logistic regression (proteins + clinical variables)</td>
<td>0.840 (0.03)</td>
<td>0.783 (0.052)</td>
<td>0.775 (0.038)</td>
<td>0.778 (0.024)</td>
<td>0.784 (0.037)</td>
</tr>
<tr>
<td>Random forest (proteins + clinical variables)</td>
<td>0.845 (0.004)</td>
<td>0.833 (0.0)</td>
<td>0.779 (0.019)</td>
<td>0.791 (0.014)</td>
<td>0.824 (0.004)</td>
</tr>
</tbody>
</table>

Table 2: The top fifteen protein biomarkers for Necrotizing enterocolitis ranked based on Random Forest variable importance method

<table>
<thead>
<tr>
<th>Rank</th>
<th>Biomarker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sex Hormone-Binding Globulin (SHBG)</td>
</tr>
<tr>
<td>2</td>
<td>Growth Hormone (GH)</td>
</tr>
<tr>
<td>3</td>
<td>von Willebrand Factor (vWF)</td>
</tr>
<tr>
<td>4</td>
<td>Cancer Antigen 125 (CA-125)</td>
</tr>
<tr>
<td>5</td>
<td>Beta-2-Microglobulin (B2M)</td>
</tr>
<tr>
<td>6</td>
<td>Tumor necrosis factor receptor 2 (TNFR2)</td>
</tr>
<tr>
<td>7</td>
<td>Interleukin-2 (IL-2)</td>
</tr>
<tr>
<td>8</td>
<td>Interleukin-6 (IL-6)</td>
</tr>
<tr>
<td>9</td>
<td>Interleukin-10 (IL-10)</td>
</tr>
<tr>
<td>10</td>
<td>Matrix Metalloproteinase-3 (MMP-3)</td>
</tr>
<tr>
<td>11</td>
<td>Macrophage Inflammatory Protein-1 beta (MIP-1 beta)</td>
</tr>
<tr>
<td>12</td>
<td>Endothelin-1 (ET-1)</td>
</tr>
<tr>
<td>13</td>
<td>Alpha-2-Macroglobulin (A2Macro)</td>
</tr>
<tr>
<td>14</td>
<td>Insulin</td>
</tr>
<tr>
<td>15</td>
<td>Prostate-Specific Antigen, Free (PSA-f)</td>
</tr>
</tbody>
</table>

References:

Quantitative Neurocognitive Phenotype of ICU Survivors: A Multimodal Data Modelling Project

David M Maslove, MD, MS, FRCPC1, Michael Wood, BA2, Stephen H. Scott, PhD2, J Gordon Boyd, MD, PhD, FRCPC1,2

1Department of Critical Care Medicine, Queen’s University, Kingston, Ontario, Canada
2Center for Neuroscience Studies, Queen’s University, Kingston, Ontario, Canada

INTRODUCTION - The last two decades have seen a significant increase in the number of patients that are surviving critical illness and injury, leading to a heightened recognition of post-intensive care syndrome (PICS). This syndrome, characterized by long term, persistent neurocognitive and neuropsychiatric dysfunction, has recently become the focus of a growing number of research programs. Important constructs remain vaguely defined, thereby undermining the precision with which predictors and outcomes can be characterized for further rigorous study. These include concepts related to the Intensive Care Unit (ICU) stay, such as delirium, as well as those related to recovery, such as good functional outcome.

Current tools to characterize these deficits include screening surveys, subjective rating scales, and neuropsychiatric tests, which tend to be cumbersome, costly, and time-consuming. At the same time, novel monitoring systems and diagnostics are increasingly capable of providing precise, objective, and granular assessments of cerebral performance and neurocognitive function, suggesting opportunities for informatics-based methods to support rigorous studies in this area. In addition to large quantities of clinical and physiologic data routinely collected in the ICU, additional modalities can provide useful data for developing and testing hypotheses in neurocritical care. One example of such a system is near-infrared spectroscopy (NIRS), which is increasingly used as a non-invasive measure of tissue oxygenation, including cerebral oxygenation.

Newer technologies allow for data-enabled monitoring and evaluation beyond the ICU, following a patient’s discharge from hospital. Functional outcomes can be more precisely quantified using a novel robotics tool (KINARM), which generates rich, quantitative metrics of upper limb motion in humans. The KINARM has been used to identify subtle neurocognitive deficits in patients after stroke, that are not detectable by routine clinical assessments. Importantly, the deficits identified correlated with quality of life after stroke. More recently, the KINARM device has been used to quantify the neurocognitive phenotype in other settings, including patients with traumatic brain injury. The KINARM battery of tests consists of 8 tasks, which together assess a variety of sensory, motor, visuospatial, and executive functions that are integral to the description of post-ICU functional status. Within each task, 6-12 performance metrics (e.g. limb speed, reaction time, accuracy) are measured. Therefore, for each patient, up to 100 performance metrics may be available. Properly analyzed, the data generated by these tools will provide a precise, quantitative neurocognitive phenotype of the ICU survivor, and identify clinical conditions in the ICU that may increase the risk of PICS.

METHODS - The Cerebral Oxygenation and Neurological Outcomes FOLLowing CriticAl Illness (CONFOCAL) research program (NCT02344043) aims to define the acute and chronic neurological complications of critical illness using highly granular data. A multimodal dataset is generated for each patient enrolled in the CONFOCAL study.

Figure 1 - Time series of high-frequency heart rate (red) and brain tissue oxygenation (BtO2, blue) signals from a critically ill patient in the intensive care unit, spanning more than 18 hours. The area shaded in grey shows a period during which concordance between heart rate and BtO2 was evident, suggesting loss of cerebral autoregulation.
merging physiologically and temporally disparate but complementary signals. Data collection begins in the ICU, where we use NIRS to measure brain tissue oxygenation (BtO₂) in critically ill patients. Delirium assessments are done using the Confusion Assessment Method for ICU (CAM-ICU), a validated screening tool for detecting delirium in critically ill patients. At 3 and 12 months following ICU discharge, we use the KINARM to assess neurocognitive function, supplemented by the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS®) evaluation.

We developed informatics tools to explore relationships between BtO₂ data, telemetry, pulse oximetry, and other physiologic data collected by the bedside ICU monitor. We merged high-frequency vital signs and BtO₂ data (Figure 1) using automated Python and R scripts, and identified correlations between physiologic parameters in the ICU, and the presence of delirium. We also used unsupervised machine learning algorithms to determine the optimal number of neurocognitive phenotypes based on raw, normalized data from the KINARM and RBANS assessments, and centroid-based clustering methods to assign patients to individual clusters. We then used clinical parameters from the ICU stay, including BtO₂ values, to classify patients into one of the outcome clusters identified, using decision trees. We determined the classification performance of the model using a leave-one-out cross-validation approach.

RESULTS - As of March, 2016, nearly half of the target enrolment of 150 patients has accrued, with 3 month outcome data available on a smaller subset of these. Early analyses have revealed important relationships between physiologic parameters that have not previously been described. Patients experiencing delirium during their ICU stay were found to have lower BtO₂ values than those who remain neurologically intact. Patients with delirium also exhibited a linear relationship between BtO₂ and hemoglobin (r=0.485, p < 0.05) which was not present in patients without delirium, suggesting a role for impaired cerebral autoregulation and oxygen delivery in the pathophysiology of delirium. Hierarchical clustering of outcome data from the first 18 patients is shown in Figure 2. The raw outcomes data partitioned into 3 clusters. The decision tree model based on clinical features at the time of ICU admission was able to classify patients into outcomes clusters with an overall accuracy of 78%.

DISCUSSION - We developed a multimodal data collection and analysis platform that merges physiologic and neurocognitive information to identify quantitative phenotypes of ICU survivorship. Future directions will involve completing enrolment of the CONFOCAL clinical trial, and further developing data analysis pipelines with a focus on time series analysis, unsupervised learning for hypothesis generation, risk factor identification, as well as more nuanced and precise gradations of functional status. These tools will support research in the area of post-ICU syndrome, an increasingly recognized clinical entity with significant morbidity and socioeconomic impact.

References

Patient use of electronic prescription refill and secure messaging, and its association with undetectable HIV viral load

Keith McInnes, ScD¹,², Stephanie Shimada, PhD¹,²,³, Amanda Midboe, PhD⁴, Kim Nazi, PhD⁵, Shibei Zhao, MPH¹, Justina Wu, MPH⁴, Casey Garvey, BSN, PhDc⁶, Thomas Houston, MD, MPH¹,³
¹Center for Healthcare Organization and Implementation Research, Edith Nourse Rogers Memorial VA Medical Center, Bedford, MA; ²Department of Health Law Policy and Management, Boston University School of Public Health, Boston, MA; ³Division of Health Informatics and Implementation Science, Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; ⁴Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, CA; ⁵Veterans and Consumers Health Informatics Office, Office of Connected Care, Veterans Health Administration, U.S. Department of Veterans Affairs, Washington, DC; ⁶School of Nursing, Bouvé College of Health Sciences, Northeastern University, Boston, MA;

Abstract

Electronic personal health records may help patients engage in their own health care. We analyzed medical record data from 3374 veterans with HIV. There was a statistically significant positive association between use of electronic prescription refill and a change from detectable to undetectable HIV viral load status between 2009 and 2012, with OR=1.36 (95% CI 1.11-1.66). Use of a personal health record may result in greater control of HIV.

Introduction

Electronic Personal Health Records (PHRs) help patients engage in management of their own health care.¹² With such systems, patients go online to request prescription refills, view laboratory results, or communicate with their providers via electronic secure messaging. The Department of Veterans Affairs (VA) PHR, known as My HealtheVet (MHV), was introduced in 2003 and has more than 3.5 million users. Veterans with HIV register for MHV at higher rates than veterans with other chronic conditions. However, it has not yet been established that use of MHV improves self-management or clinical outcomes for veterans with HIV. The present study investigates whether there are associations between use of MHV, and specifically its prescription refill (Rx Refill) and secure messaging (SM) functions, and HIV viral load in US veterans. Viral load measures how well HIV medications are working for a person infected with HIV, with the goal being a viral load so low that it is considered undetectable.³

Methods

We conducted a retrospective cohort study, analyzing medical record data from a cohort of 3374 veterans with HIV who received VA care between 2009 and 2012. Multivariable analyses accounted for facility level clustering and examined whether the dichotomous outcome of viral load status (detectable or undetectable) was associated with use of either Rx Refill or SM. Analyses controlled for age, gender, marital status, race/ethnicity, urban/rural residence, economic need, and comorbidities.
Results

There was a statistically significant positive association between use of Rx Refill and a change from detectable to undetectable HIV viral load status between 2009 and 2012, with OR=1.36 (95% CI 1.11-1.66). There was a similar association between SM use and change in viral load status with OR=1.28 (95% CI 0.89-1.85), but without achieving statistical significance. Analyses did not demonstrate a dose-response of Rx Refill or SM use on the achievement of undetectable viral load.

Conclusions

These findings suggest that use of a PHR among US veterans with HIV may result in greater control of this chronic illness, as measured by undetectable viral load. Exactly how this happens is unclear, but it is possible that PHR features may help increase patient-provider communication or patient motivation to self-manage one’s health. Alternatively PHRs may contribute to improved medication adherence by reducing gaps in patient supply of HIV medications on hand. This study’s limitations include the lack of experimental design, and unmeasured confounders such as health literacy and computer literacy. Due to these limitations it is possible that PHR use does not improve health processes or health measures such as viral load, but rather is a marker for high levels of engagement in one’s health. Additional analyses, and if possible including experimental designs, are needed to better understand the mechanism by which PHR use may (or may not) be contributing to health care processes and outcomes.

References

Feature Portability in Cross-domain Clinical Coreference

Timothy A. Miller, PhD1, Dmitriy Dligach, PhD1, Chen Lin1, Steven Bethard, Phd2, Guergana K. Savova, PhD1

1Boston Children’s Hospital and Harvard Medical School, Boston, MA; 2University of Alabama-Birmingham, Birmingham, AL

Introduction

Clinical coreference resolution is the task of grouping together multiple entity mentions in a text that correspond to the same real word entity. This has been a task of interest in natural language processing for decades, but has proven difficult and is still considered unsolved. Even to the extent that coreference systems can be built for clinical text, and score on par with systems in the general domain, there is some concern that these trained models are overfit to the domain of the training data. In this work we focus on the overfitting problem, and take advantage of new gold standard coreference datasets to explore the extent to which systems trained on one corpus can apply to another. We focus especially on mention-specific features, looking at the relative portability of word identity features, part of speech tag features, and dependency-tree-based features.

Methods

Datasets: We use two clinical datasets containing gold standard coreference annotations. The first is THYME (Temporal History of Your Medical Events), a temporally-annotated corpus that also contains coreference annotations on a subset of notes1. We use the THYME training data as our supervised training source, containing 273 colon cancer notes, including clinical notes and pathology notes that are intended to cover the disease course from initial visit to diagnosis to treatment. The second corpus we use is from the cancer deep phenotyping project (DeepPhe)2, which focuses on extracting detailed descriptions of cancer phenotypes from a range of different cancer types. From this corpus we use a preliminary version of the training data, containing 48 breast cancer notes, including radiology, discharge summaries, progress notes, and others. This corpus is not yet publicly available.

Methods: Coreference resolution is typically modeled as a supervised machine learning problem. One of the most common and successful approaches is the pairwise entity paradigm3, where the classifier is trained to decide whether pairs of mentions refer to the same entity. The outcomes from this classifier are then passed to a second phase where any inconsistent labels must be reconciled. We focus on a newer paradigm, the mention synchronous approach4, in which coreference chains are built incrementally, and newly encountered mentions are compared against existing chains in the document. This is part of a broader research agenda to track clinical events across notes, and so we model the known information about an event (e.g., tumor size, location) as it is mentioned and potentially use this information to resolve the notoriously difficult cross-document coreferent mentions. Evaluation in coreference resolution is performed using several different metrics (MUC, B^3, and CEAF), each with strengths and weaknesses, and recent work has used a combined score that averages the three (CoNLL – Computational Natural Language Learning5). We use the official CoNLL scoring tool, version 8.01.6

One difficult aspect of clinical coreference resolution is that clinical notes can be quite long and contain many entities. Since making every possible comparison is computationally prohibitive, we use a number of heuristics to limit the number of comparisons. First, we attempt to link a mention to any cluster that has a mention in the last five sentences. Next, we attempt to link to any cluster with an element that is part of a section header. We also allow a possible link to any cluster with more than one element. Finally, we allow linking to any cluster where a mention has the same headword as the current mention.

Our system uses a variety of standard features – adapting strong features from our previous pairwise system7 as well as newer features that look at the likelihood of a mention being a singleton8, the selectional preferences of different UMLS (Unified Medical Language System) semantic types, and that take advantage of the fact that the antecedent is a cluster instead of a single mention. Our system uses a linear support vector machine (Liblinear) to do classification.

In this work we examine the value of token features for coreference, since these features are very strong in other relation extraction tasks. Our experiments on THYME development set found that performance without token features was 56.8 (CoNLL), while with word token features it was 58.4. Given the difficult task, on a system we had
spent quite a bit of time developing, this was a surprisingly large gain. But with the specificity of our training domain, we had some concern that features focused on word identity would not generalize. We thus explore the potential gains possible with different representations of token-level features, with the explicit aim of evaluating on a new corpus which we had not seen. We looked at several configurations: Baseline (no token features) Word (word identity in three words preceding each mention, inside each mention, and three words following each mention), POS (part of speech tags using the same context as word features), and DEP (combinations of the mention head word and POS tag with its dependency tree parent head word and POS tag).

We hypothesize that while word features have a positive effect when trained and evaluated on THYME, they would have no effect or harmful effect in the new domain. We expect POS features to generalize better, and we expect dependency features to be better yet, since they rely on syntactic context, which is less reliant on local variation.

### Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>Baseline</th>
<th>Word</th>
<th>POS</th>
<th>DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>THYME (train)</td>
<td>THYME (dev)</td>
<td>56.81</td>
<td>58.42</td>
<td>57.45</td>
<td>58.0</td>
</tr>
<tr>
<td>THYME (train)</td>
<td>DeepPhe (train)</td>
<td>33.20</td>
<td>31.37</td>
<td>32.32</td>
<td>31.36</td>
</tr>
</tbody>
</table>

### Discussion

First, we should note that cross-domain performance is probably not as bad as it appears, because the two corpora differ in representation of de-identified names, which are quite common, and our system is tuned for THYME. THYME de-identification creates more realistic-looking names, so on real data we would thus expect the system to perform reasonably well on names. Our results (see table above) show that word features were indeed not valuable at porting to a new domain, and seem to hurt performance. We were surprised, however, to see that the same is true for POS features and DEP features. While DEP features combine Word and POS features with syntactic context, they too proved to be a drain on cross-corpus performance. These results could be taken two ways; first, they could be used to suggest that token identity features should be omitted for more robust coreference resolution. Alternatively, these results could be used to suggest that new methods that allow systems to incorporate target-domain lexicon into the feature representation would improve coreference system portability.

### References

Applying Active Learning to Clinical Abbreviation Disambiguation in Real Time

Sungrim Moon, Ph.D.\(^1\), Yukun Chen Ph.D.\(^3\), Jingqi Wang M.S.\(^1\), Joshua C. Denny M.D., M.S.\(^4,5\), S. Trent Rosenbloom M.D., MPH\(^4,5\), Ky Nguyen R.N., B.S.N.\(^1\), Toluulola Dawodu R.N., B.S.N.\(^2\), Hua Xu Ph.D.\(^1,3\)

\(^1\)School of Biomedical Informatics, \(^2\)School of Nursing, The University of Texas Health Science Center at Houston, Houston, TX, USA; \(^3\)Pieces Technologies Inc, Dallas, TX, USA; \(^4\)Department of Biomedical Informatics, \(^5\)Department of Medicine, Vanderbilt University, Nashville, TN, USA

**Objective:** This study aims to assess the effectiveness of active learning\(^4\) in a real-time annotation process for building machine learning based classifiers for clinical abbreviation disambiguation. Previous studies\(^1\) have shown that active learning could achieve significant reduction in the size of annotated data and annotation time primarily in the simulation mode to achieve desired model quality. However, very few studies have also considered the actual annotation cost in active learning in a real-time setting, especially in the clinical domain. In this study, we used an existing active learning annotation interface, DUALIST\(^5\), to build real-time clinical abbreviation disambiguation. We designed a user study with both physicians and nurses as annotators to investigate the efficiency of active learning on the two types of experts.

**Methods:** In the current study, we built classifiers for six ambiguous clinical abbreviations: CA, CC, DM, LAD, PE, and RA. Two physicians and two nurses were asked to annotate sentences containing ambiguous clinical abbreviations, via either active learning or passive learning (random sampling) in a fixed time period of five minutes. DUALIST implements an uncertainty-based querying algorithm integrated with multinomial Naïve Bayes as the classifier for active learning. We evaluated the performance of the classifiers from active learning or passive learning using an independently annotated dataset of ambiguous abbreviations. We then generated learning curves, which plots the accuracies of the models vs. annotation time, as well as the area under the learning curve (ALC), which represents a global score of the learning process for each annotator, disambiguation task and sampling algorithm. We also report the annotation speed (the number of sentences annotated per minute) and the annotation difficulty (the ratio of sentences annotated as unsure and errors over the total annotated sentences).

**Results:** Overall, active learning outperformed passive learning in terms of the final accuracy on the test set. Table 1 shows results from all twelve experiments for the physicians and nine out of twelve for the nurses; the average final accuracies were 0.87 and 0.85 in the active learning and passive learning modes, respectively. A comparison of the effectiveness between physicians and nurses shows that ALC scores from active learning for physicians were higher (3.2% improvement on average from active learning to passive learning) than nurses (1.7% deterioration on average). Table 2 shows that annotations in active learning experiments had higher annotation difficulty (6.8% and 15.5% for physicians and nurses, respectively) than passive learning (1.1% for both groups in annotation difficulty). In other words, annotators had higher errors and uncertainty to annotate senses in active learning. Furthermore, active learning had lower annotation speed than passive learning. Also, physicians annotated more sentences in relatively short times as compared to nurses. Figure 1 presents the learning curves for the abbreviation, RA, for one physician (P1) and one nurse (N2). To achieve an accuracy of 0.95, active learning decreased annotation time by 77% (3.8 reduced minutes) and annotated sample size by 87% (101 reduced samples) than passive learning. In fact, passive learning could not reach 0.95 accuracy in five minutes for the same task. Figure 1 also shows that physician annotated 41 more sentences with higher accuracy (0.07) than nurse as well.

**Discussion:** Active learning showed higher performance of both final accuracies and ALC scores than passive learning for the real-time clinical abbreviation disambiguation tasks. We also found that active learning could be more beneficial to advanced clinical experts because physicians annotated more informative cases with lower error rate and higher speed than nurses. Despite the relative simplicity of this study, it presented the applicability of the active learning to build real-time WSD classifiers in the clinical domain.

**Acknowledgement:** This study was supported by grant from the NLM R01LM010681 and Mayo Clinic.
Table 1. Final accuracy with ALC score per strategy on the test set

<table>
<thead>
<tr>
<th>Abbrev</th>
<th>Final accuracy (AL/PL)</th>
<th>ALC Score (AL/PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>CA</td>
<td>0.85/0.78</td>
<td>0.85/0.84</td>
</tr>
<tr>
<td>CC</td>
<td>0.81/0.80</td>
<td>0.81/0.79</td>
</tr>
<tr>
<td>DM</td>
<td>1.00/1.00</td>
<td>1.00/0.97</td>
</tr>
<tr>
<td>LAD</td>
<td>0.99/0.98</td>
<td>0.99/0.96</td>
</tr>
<tr>
<td>PE</td>
<td>0.76/0.75</td>
<td>0.73/0.69</td>
</tr>
<tr>
<td>RA</td>
<td>0.99/0.91</td>
<td>0.99/0.94</td>
</tr>
</tbody>
</table>

P1 = Physician 1, P2 = Physician 2, N1 = Nurse 1, N2 = Nurse 2, ALC = Area under the Learning Curve, AL = Active learning mode, and PL = Passive learning mode

Table 2. Annotation speed and difficulty per strategy on the test set

<table>
<thead>
<tr>
<th>Abbrev</th>
<th>Annotation speed (sentences) (AL/PL)</th>
<th>Annotation difficulty (percentage %) (AL/PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>CA</td>
<td>15.6/24.6</td>
<td>12.6/26.0</td>
</tr>
<tr>
<td>CC</td>
<td>16.2/21.6</td>
<td>7.8/16.6</td>
</tr>
<tr>
<td>DM</td>
<td>29.2/37.6</td>
<td>22.6/21.0</td>
</tr>
<tr>
<td>LAD</td>
<td>18.2/24.4</td>
<td>14.4/19.6</td>
</tr>
<tr>
<td>PE</td>
<td>17.6/22.6</td>
<td>11.0/16.4</td>
</tr>
<tr>
<td>RA</td>
<td>22.2/23.2</td>
<td>20.4/17.6</td>
</tr>
</tbody>
</table>

P1 = Physician 1, P2 = Physician 2, N1 = Nurse 1, N2 = Nurse 2, AL = Active learning mode, and PL = Passive learning mode

Figure 1. Learning curves that plot accuracy of a WSD classifier vs. annotation time for RA abbreviation

Reference

Towards Phenotyping Stroke: Leveraging Electronic Health Record Data to Identify Stroke Cases in a Large-scale Epidemiological Study

Yizhao Ni¹, Charles Moomaw², Kathleen Alwell², Dawn Kleindorfer², Daniel Woo², Opeola Adeoye², Matthew Flaherty², Simona Ferioli³, Jason Mackey³, Felipe De Los Rios La Rosa², Sharyl Martini³, Pooja Khatri², Brett M. Kissela²

¹Cincinnati Children’s Hospital Medical Center, ²University of Cincinnati, Cincinnati, OH ³Indiana University, Indianapolis, IN, ⁴Michael E. DeBakey VA Medical Center, Houston, TX

Introduction

Administrative epidemiological studies typically use International Classification of Diseases (ICD) codes or death certificate data to identify stroke cases. However, they are limited by positive predictive values (PPVs) ranging from 6% to 97% across study designs and stroke types. A paucity of studies utilizes physician chart review of electronic health record (EHR) data to confirm cases and has greater accuracy, but these studies are costly and labor intensive. The large volume of false positives has hindered the re-use of collected samples for research requiring high statistical power, such as the Electronic Medical Records and Genomics (eMERGE) studies, which use patient cohorts of high purity to identify genome-wide associations. To minimize loss of statistical power, we investigated machine learning methods to identify high-precise stroke cohorts based on a broad array of EHR data. Our objective was to develop a machine learning-based approach that identifies stroke cases with high predictive performance.

Data and Method

We utilized stroke events collected in the Greater Cincinnati/Northern Kentucky Stroke Study, which is a large-scale, population-based epidemiological study that measures temporal trends in stroke incidence rates in a population of 1.3 million. The study identified hospitalization events with stroke-related diagnoses (primary/secondary ICD9 codes 430-438) from 17 regional hospitals in 2005 and 2010. Detailed information from patients’ EHRs was abstracted for each potential event by trained research nurses. The abstracts contained 281 structured variables categorized into 16 sets, which are presented in Table 1. Each abstract was reviewed by at least one stroke physician to determine if the event was a stroke case and, if so, the stroke subtype. The study personnel reviewed a total of 8131 potential events, of which 6987 samples were adjudicated to be stroke cases.

The event labels adjudicated by the physicians served as the gold standard to train and evaluate the predictive models. The baseline algorithm used ICD9 codes (430-431, 434-436) that identified stroke cohorts with high precision. We then compared the baseline with four machine learning classifiers: logistic regression (LR), support vector machines with polynomial (SVM-P) and radial basis function (SVM-R) kernels, random forest (RF), and three-layer artificial neural network (ANN) with L2-norm regularization. We chose these classifiers to allow for the possibility of linear and non-linear relationships between the variables and event labels. A nested ten-fold cross-validation was used to optimize model parameters and compare model performances. The model parameters (i.e., regularization parameters for LR, SVM-P, SVM-R and ANN; degree parameter for SVM-P; γ parameter for SVM-R; and number of trees for RF) were optimized using grid search parameterization. A core set of variables (DEMO, VI, ICD9, SS and CT/MRI) was selected based on the physicians’ expert knowledge. The nominal variables (e.g. ICD9 codes) were converted to binary features, whereas the numerical variables (e.g. blood pressure) were discretized into bins before converting to binary features. Macro-average PPV, sensitivity (SENS), negative predictive value (NPV), specificity (SPC) and the area under the ROC curve (AUC) were calculated to assess model performance. To explore all variable sets and validate their contributions, we also used the best first search strategy to incrementally add the variables into a LR model and assessed the corresponding performance.

Results and Discussion

The performances of the predictive models are presented in Table 2. Compared with the baseline, all machine learning algorithms yielded significantly better AUCs (p<1.0E-10 under paired t-test). The RF achieved the best AUC (89.51%), and its improvements over the other models were statistically significant. Figure 1 shows the false discovery rate (1-PPV)-sensitivity curves achieved by the models. We observed that the ICD9-coded predictor could not achieve an acceptable PPV (95%) to support studies requiring high precision such as the eMERGE studies. On the contrary, the RF could provide approximately 90% of the true cases for these studies under an acceptable PPV of 95% (dotted line, Fig. 1). Figure 2 shows the AUC curve when incrementally adding the variable sets using best first search. The top predictive variables included signs and symptoms, CT/MRI findings, ICD9 codes, interventions, demographics, visit information, and ED assessments. The core set designated by expert knowledge covered most of the predictive variables and its performance (85.90%) was close to the best AUC ever achieved (85.95%, Fig. 2).
The developed models identified stroke cases with high precision and adequate sensitivity, allowing high statistical power for retrospective stroke research without labor-intensive clinician review. The predictive variables identified in the study could be used to assist physician chart review and to develop generalizable stroke phenotype algorithms.

Table 1. Description of the variable categories used in the study.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMO</td>
<td>Patient demographics including age, gender, race, ethnicity and insurance type</td>
</tr>
<tr>
<td>MH</td>
<td>General medical history prior to hospitalization (e.g., history of hypertension Yes/No)</td>
</tr>
<tr>
<td>SH</td>
<td>History of stroke prior to hospitalization (e.g., ischemic stroke Yes/No)</td>
</tr>
<tr>
<td>SA</td>
<td>Patients’ history of substance use (smoking, alcohol and street drugs)</td>
</tr>
<tr>
<td>VI</td>
<td>Visit information at time of admission (e.g., residence at home/nursing home/assisted living)</td>
</tr>
<tr>
<td>ICD9</td>
<td>Primary and secondary ICD-9 codes on patients’ discharge lists</td>
</tr>
<tr>
<td>ED</td>
<td>The evaluations (e.g., blood pressure and Glasgow Coma Score) performed in EDs</td>
</tr>
<tr>
<td>SS</td>
<td>Signs and symptoms during hospitalization (e.g., weakness, headache, vision and speech)</td>
</tr>
<tr>
<td>CX</td>
<td>Complications and new diagnoses during hospitalization (e.g., pain, seizure and brain edema)</td>
</tr>
<tr>
<td>SE</td>
<td>The stroke-related evaluations (e.g., NIH stroke scale done by stroke team? Yes/No)</td>
</tr>
<tr>
<td>CT/MRI</td>
<td>If CT &amp; MRI was performed and, if so, the findings (e.g., normal and infarct)</td>
</tr>
<tr>
<td>ANG</td>
<td>If MR/CT/Cerebral angiography was performed and, if so, the findings (e.g., normal/abnormal)</td>
</tr>
<tr>
<td>CU</td>
<td>If carotid ultrasound was performed and, if so, the findings (e.g., normal/abnormal)</td>
</tr>
<tr>
<td>ECHO</td>
<td>If echocardiogram was performed and, if so, the findings (e.g., cardiomyopathy Yes/No)</td>
</tr>
<tr>
<td>TT</td>
<td>The interventions performed (e.g., aneurysm clipping/coiling and clot evacuation Yes/No)</td>
</tr>
<tr>
<td>TH</td>
<td>The therapies performed (e.g., physical, occupational and speech therapy Yes/No)</td>
</tr>
</tbody>
</table>

*Due to the limited space, we defer the description of all variables to an extended manuscript.

Table 2. Classification performance of the predictive models.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Ten-fold Nested Cross Validation [%]</th>
<th>PPV</th>
<th>SENS</th>
<th>NPV</th>
<th>SPC</th>
<th>AUC</th>
<th>p value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>92.26</td>
<td>93.62</td>
<td>57.16</td>
<td>52.01</td>
<td>87.24</td>
<td>7.70E-4</td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>89.28</td>
<td>97.04</td>
<td>61.73</td>
<td>28.85</td>
<td>85.90</td>
<td>4.88E-6</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>93.61</td>
<td>92.39</td>
<td>57.28</td>
<td>61.40</td>
<td>89.51</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>SVM-P</td>
<td>90.73</td>
<td>96.98</td>
<td>68.80</td>
<td>39.44</td>
<td>87.25</td>
<td>6.65E-5</td>
<td></td>
</tr>
<tr>
<td>SVM-R</td>
<td>90.89</td>
<td>96.64</td>
<td>66.80</td>
<td>40.85</td>
<td>87.63</td>
<td>3.74E-4</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>88.11</td>
<td>62.54</td>
<td>17.45</td>
<td>48.41</td>
<td>55.48</td>
<td>3.98E-10</td>
<td></td>
</tr>
</tbody>
</table>

*The p value was calculated by comparing the AUC between the best algorithm (RF) and the other algorithms using paired t-test.

†N/A indicates that the performances between the two algorithms are identical and no p value is returned.

Figure 1. The false discovery rate-sensitivity curves.  
Figure 2. The AUC curve when adding the variables.

References

Evaluation of Whether Radiology Decision Support Embedded in the Order Entry Function of an Electronic Medical Record System Improves Appropriateness of Cross-Sectional Imaging Orders

Ted E. Palen, PhD, MD, MSPH1 and Susan Shetterly, MS2
1Institute for Health Research, Colorado Permanente Medical Group, Denver, Colorado
2Institute for Health Research, Denver, Colorado

Abstract
Possibly 30% of diagnostic imaging may be unnecessary or inappropriate. We evaluated whether the use of radiology decision support improved the appropriateness of radiology orders. The proportion of appropriate orders during baseline was 78.6% compared to 80.6% after the appropriateness alerts were turned on ($p < 0.06$). Among primary care physicians the adjusted rate of appropriate imaging orders was 77.8% at baseline and 81.0% after the alert was “on” ($p = 0.005$).

Introduction
The use of diagnostic imaging has rapidly increased over the last two decades.1-3 Studies indicate that perhaps 26% to 30% of diagnostic imaging procedures may be either unnecessary or inappropriate.1,4 Not only does increased use of medical imaging lead to increases in health care expenditures it also leads to patients being exposed to increasing amounts of ionizing radiation.2,4 Investigating appropriateness of medical practice is a logical extension of the empiric research which has revealed substantial variation in the type and intensity of health services, including medical imaging services.4 The American College of Radiology (ACR) has developed Appropriateness Criteria (AC) for imaging studies based on clinical indications.5 The ACR AC ratings range from 1 (least) to 9 (most) appropriate study for a clinical indication. It also displays the relative level of radiation exposure for specific imaging studies. In an attempt to address the rising rate of imaging and the concomitant rise in unnecessary or inappropriate imaging studies, radiology decision support tools are being incorporated into electronic medical records (EMR) which provides clinicians with evidence-based decision support at the time of ordering imaging studies. In comments submitted by Dr. Fridsma (President of AMIA) and Dr. Payne (AMIA Board Chair) to the Centers for Medicare & Medicaid Services (CMS), they stated that AMIA believes CMS has an unprecedented opportunity to learn which components of these legacy programs [payment systems] will effectively support our healthcare system in moving toward the triple aim and recommended that CMS engage medical informatics expertise more broadly to understand how technology should be leveraged to improve care experience, expense and efficacy.5 Conference participants attending this presentation will be presented with results regarding the use of radiology ordering decision support at the point-of-care. Studies like this will help inform the objectives of both AMIA and CMS in understanding how to incorporate technology into care processes.

Objective
Our goal was to evaluate whether the use of radiology decision support embedded in an EMR point-of-care ordering system will improve the appropriateness orders for cross-sectional radiology studies.

Study Site
The study site was an integrated health care delivery system. It serves a patient population of over 600,000 members. The health plan employs over 6,000 staff and the medical group consists of over 1,000 physicians. The system deployed a commercial EMR in 2004. The study involved a step-wedge, cluster design, preceded by a “run-in” period. Twenty-four ambulatory clinic locations were randomly divided into four clusters. Radiology decision support software using the ACR AC criteria was interfaced with the ambulatory electronic medical record order entry system. During the “run-in” period, users were required to enter either a coded and or free-text clinical indication at the time of ordering an imaging study. However, during this period the decision support software did not display the AC rating for the imaging study ordered, although the software calculated and stored the rating. After a six week “run-in” period the ACR AC rating alert was activated for users in cluster one clinics who had selected one or more coded clinical indication when ordering cross-sectional imaging (CT or MRI) studies. At 2-4 week intervals the ACR AC rating alert was activated for users in cluster two, three, and four clinics. If the AC rating is in the poor or not indicated range (ACR score 1-6) the software suggests a more appropriate imaging study. The ordering clinician can then either
precede with the original order (indicating an over-ride reason), choose the suggested imaging study, or cancel the order completely.

Results

After close of the study period we analyzed the data using generalized linear mixed models with a logit link (e.g. logistic regression) where the AC rating was treated as a binary outcome variable (AC rating 1-6 inappropriate and 7-9 appropriate). After excluding records that did not have an AC rating, the final sample included 18,121 (58.9%) of orders. The orders were made by 941 different clinicians (71.8% MD, 13.8% PA, 7.6% NP, 6.0% DO, and 0.9% OD, DPM, or DC). The patient median age was 56.7 (SE 16.3) and 56.8% female. During the period when the alert was “off” the adjusted rate of order appropriateness (AC rating of 7-9) was 78.6% compared to 80.6% after the AC rating alert was “on” (p = of 0.06). Among primary care physicians the adjusted rate of appropriate imaging orders was 77.8% at baseline and 81.0% after the alert was “on” (p = 0.005).

Discussion

Even though we saw an overall increase in the proportion of appropriate imaging studies it was not statically significant. However, we are uncertain whether this increase maybe clinically significant. That is, we do not know if more appropriate ordering improved clinical outcomes or decreased the need for additional imaging studies. We were surprised at the high proportion of appropriate imaging studies at baseline (“run-in” period). This may indicate a computer Hawthorne effect, that is, perhaps just requiring clinicians to enter a coded clinical indication may influence their clinical decision making. Or, it could indicate that baseline ordering behavior is highly evidence-based due to other factors. These questions all deserve additional investigation.

Conclusion

We used decision support software embedded into the order entry system of an electronic medical record system. The software displayed the ACR AC rating for the appropriateness of the cross-sectional imaging study ordered based on the indication selected by the clinician. The proportion of appropriate imaging studies ordered for specific clinical indications increased. The overall rate increase was not statistically significant. However, for the sub-analysis of primary care physicians only, we did see a statistically significant increase in the rate of appropriate ordering.

References

4. Lehnert BE, Bree RL. Analysis of appropriateness of outpatient CT and MRI referred from primary care clinics at an academic medical center: how critical is the need for improved decision support? J Am Coll Radiol 2010;7:192-197.
Detailed Clinical Models for Nursing Actions

Hyeoun-Ae Park, PhD1, Joo Yun Lee, MS1
1Seoul National University, Seoul, Republic of Korea

Abstract
The aims of this study were to develop and evaluate detailed clinical models for nursing actions in perinatal care. We extracted entities (dyads of action and target concepts) of nursing actions, attributes and possible values needed to describe either action or target concepts in more detail by analyzing nursing documentations, reviewing the literature, and interviewing nurse experts. In total, 233 models were developed with 34 action, 212 target, 211 attributes, and 534 value concepts.

Introduction
The potential benefits of electronic health records (EHRs) include improvements in patient safety and quality of care, reduced medical errors, reduced cost and improved ability to conduct research and population health1. However, these benefits depend on the semantic interoperability of data to be shared and exchanged.

One way to ensure semantic interoperability is to have an agreed clinical data structure2. Agreed data structures may be referred to under many different names, such as detailed clinical models (DCMs), clinical element models, care information models, clinical content models, clinical templates, archetypes, clinical fragments, and more3. A DCM is a relatively small, standalone information model designed to express a clinical concept in a standardized and reusable manner. A DCM consists of an entity-attribute-value (EAV) triplet where an entity is a core or focus concept of a datum, an attribute is a qualifier that represents the entity in more detail, and a value set is a uniquely identifiable set of valid values describing instances of an attribute.

While the importance of developing DCMs for nursing has been addressed by many nurse informaticists4, there are only a few reports on this topic in nursing literature. To the best of our knowledge, these efforts mainly focused on the nursing problems and/or nursing diagnoses5,6. The aims of this study were to develop and evaluate detailed clinical models for nursing actions and or interventions in perinatal care.

Methods
First, we extracted action and target concepts which are the core concepts of nursing action, attributes and possible values sets needed to describe either action or target concepts in more detail by analyzing narrative nursing action statements, and reviewing related literature. For this, we analyzed 27,596 narrative statements used to describe nursing services, actions and/or interventions in electronic nursing records of 63 women who were hospitalized to give a birth from March 1 to April 30, 2012 at a tertiary teaching hospital in Korea. We also analyzed 795 unique narrative nursing activity statements in 26 NIC nursing interventions relevant to perinatal care based on SCOG guideline. In addition, we reviewed text books on perinatal care, AWHONN'S perinatal nursing and ISO's Conceptual framework for patient findings and problems in terminologies to identify additional action and target concepts, attributes and possible value sets.

Second, DCMs were modeled by creating dyads of action and target concepts as entities and linking these entities with possible attributes with value sets and optionalities.

Third, DCMs were validated by eight nurse experts on the reusability, non-ambiguity, relevancy, exclusiveness, completeness, and consistency using 4-point rating scale (from 1 = strongly disagreed to 4 = strongly agreed). The coverage of the DCMs was evaluated by mapping concepts extracted from 1,092 narrative statements used to describe nursing services, actions and/or interventions of 5 women who were hospitalized to give a birth from May 1 to 51, 2012 at a the same tertiary teaching hospital to the EAV triplet concepts.

Results
In total, 233 DCMs were identified from analysis of narrative nursing action statements from nursing documentation, and literature review. Three more attributes and 7 more value sets were added after consulting the nurse experts. These 233 DCMs consists of 233 entities (dyads of 34 action and 212 target concepts) with 214 attributes, and 541 value sets. There were 1 to 33 target concepts linked to one action concept with an average 6.7 target concepts. There were 1 to 3 action concepts linked to one target concepts. Action concepts such as ‘applying’, ‘assessing’ and ‘inquiring’ has the most number of targets with 33. Target concept linked with the most number of action concepts was ‘blood product’. Each entity linked to 1 to 9 attributes with ‘administering
intravenous infusion’ with the most number of attributes. Table 1 presents the ‘teaching breastfeeding’ DCM as an example.

**Table 1. Detailed clinical model for ‘teaching breastfeeding’**

<table>
<thead>
<tr>
<th>Entity</th>
<th>Attribute</th>
<th>Optionality</th>
<th>Data type</th>
<th>Value set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching breastfeeding (10019502)</td>
<td>Contents of Teaching breastfeeding</td>
<td>mandatory</td>
<td>SC</td>
<td>Benefits / Contraindications / Anatomy and physiology of lactation / Breastfeeding process: techniques / Methods of expression, storage, and alternate feeding of breast milk</td>
</tr>
<tr>
<td>Teaching (10003645 Breastfeeding)</td>
<td>Recipient</td>
<td>optional</td>
<td>SC</td>
<td>Patient / Care giver</td>
</tr>
<tr>
<td></td>
<td>Teaching material</td>
<td>optional</td>
<td>SC</td>
<td>Written material / Audio-visual</td>
</tr>
<tr>
<td></td>
<td>Teaching method</td>
<td>mandatory</td>
<td>SC</td>
<td>Lecture / Demonstration / Collaboration</td>
</tr>
<tr>
<td></td>
<td>Teaching place</td>
<td>optional</td>
<td>SC</td>
<td>Lactation room / Patient's room / Nurse’s station / Lecture room</td>
</tr>
<tr>
<td></td>
<td>Delivery formats</td>
<td>optional</td>
<td>SC</td>
<td>Individual / Group</td>
</tr>
<tr>
<td></td>
<td>Duration</td>
<td>optional</td>
<td>REAL</td>
<td>Unit: min</td>
</tr>
</tbody>
</table>

* SC : Character String with Code / REAL : Real Number

Rating scores for reusability, non-ambiguity, relevancy, exclusiveness, completeness, and consistency were 0.91, 0.96, 0.87, 0.96, 0.80, 0.96 and 0.84 respectively. Based on the experts’ evaluation, we deleted 5 entities, and 6 attributes; we added 7 entities, 23 attributes and 6 value sets; and we revised 3 entities, 1 attribute and 2 value sets.

For coverage evaluation, 192 unique statements extracted from 1092 narrative nursing action statements were mapped to DCMs. As presented in Figure 1, 174 statements were fully mapped, 7 partially mapped and 11 not mapped. Eleven not mapped statements need new action-target dyads. Six out of seven partially mapped statements need new attributes and remaining one needs new value.

**Figure 1. Mapping narrative nursing statements to detailed clinical models components**

**Conclusion**

DCMs developed in this study will be used to generate nursing narratives using natural language generation technology. Narratives generated from the DCM will guide a structured and standardized data entry and ensure semantic interoperability of data to be shared and exchanged.

**Acknowledgement**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2010-0028631).

**References**

A probabilistic model for learning relationships between diagnosis codes and clinical free text

Adler Perotte, MD, MA1, Noémie Elhadad, PhD1
1Department of Biomedical Informatics, Columbia University, New York, NY

Introduction
In this work, we aim to develop a probabilistic model that simultaneously models the hierarchical nature of a diagnosis code vocabulary and captures the relationships between diagnosis codes—including non-billing codes—and clinical free text. We focus on diagnosis codes from the International Classification of Diseases, Ninth Edition (ICD-9). Effective representations of these relationships can enable several applications including (i) the ability to highlight terms of interest in documents associated with a given diagnosis code; (ii) the ability to search for billing codes associated with a given free-text query; and (iii) the ability to search for documents of interest based on billing or non-billing diagnosis codes.

Several models and systems have been developed that leverage natural language processing (NLP) to process clinical documents, such as discharge summaries, and assign billing codes associated with that episode.1,2,3 In contrast, our goal with this is to create a flexible representation of the relationships between terms and diagnosis codes that can serve as the foundation for many applications including billing code assignment.

Our approach is based on a novel probabilistic model called the Hierarchically Identified Latent Dirichlet Algorithm (HILDA) and builds on prior work in statistical NLP to model diagnosis codes and document text simultaneously while respecting the hierarchical constraints of the ICD taxonomy. As we will demonstrate, the parameters of this model are interpretable and useful.

Methods
HILDA is a model for bag-of-word data with multiple hierarchical labels. This model is an extension of Latent Dirichlet Allocation4 (LDA) and closely related to Labelled LDA and Hierarchically Supervised LDA.3 This work differs by creating a one-to-one relationship between diagnosis codes and topics while also respecting the hierarchy of the vocabulary. In particular, the observation of a leaf diagnosis code implies the observation of ancestor diagnosis codes. Of note for this work, unobserved diagnosis codes that are not the ancestor of an observed code are considered negative.

Model Description
Let \( \mathbf{w}_d = \{ \mathbf{w}_{1,d}, \ldots, \mathbf{w}_{N_d,d} \} \) be the set of \( N_d \) terms in document \( d \).
Let \( y_{r,d} \) be the root of the ICD-9 hierarchy for document, \( d \), and \( y_{a,d} \) represent the children of the root where \( a = \{1, 2, \ldots, A\} \) and \( A \) is the total number of direct children of the root. Each subscript for the set of variables associated with the ICD-9 hierarchy represent a level. Therefore, codes that are 6 levels deep will have 6 subscripts indicating the branch at each level in addition to the subscript representing the document, \( d \). Let \( z_{n,d} \) represent the topic/diagnosis code assignment for a given term, \( \theta_d \) represent the distribution over topics/diagnosis codes for the document, and \( \phi_c \) represent the distribution over terms associated with each topic/diagnosis code.

A collapsed Gibbs sampler was developed to perform inference in this model, integrating out \( \theta_d \) and \( \phi_c \). This model was trained on a corpus from the MIMIC-II database of 20533 clinical documents (discharge summaries) and associated ICD-9 codes. There was an average of 9.24 codes and 606.27 words per document.

Results
Although predictive performance was not the primary goal of this work, HILDA predicts diagnosis codes in a held-out set with an F1-score of 37.7%. This is competitive with other discriminative classifiers, including a hierarchical
SVM. Related work predicts using random vector representations. In that work, at least one diagnosis code was in the top 10 predictions in 20% of cases whereas we observe a performance of 48%. However, this may not be the most informative performance metric and the corpus studied in that work is not very comparable to MIMIC-II.

Table 1. Word-based representations for a family of diagnosis codes. Non-billing codes are the diagnosis codes that are not leaves of the ICD taxonomy and are not used for billing.

<table>
<thead>
<tr>
<th>Diagnosis Code</th>
<th>Malignant neoplasm of lymphatic and hematopoietic tissue (200-208) – Non-billing</th>
<th>Myeloid leukemia (205) – Non-billing</th>
<th>Acute myeloid leukemia (205.0) - Billing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated Terms (top 15)</td>
<td>mg, chemotherapy, marrow, bone, showed, started, continued, ct, biopsy, transplant, bmt, negative, acyclovir, received, cefepime</td>
<td>blood, platelets, mds, myelodysplastic, hydroxyurea, percent, syndrome, platelet, oncology, danazol, hematology, myelogenous, expired, transfusion, today</td>
<td>aml, po, blasts, induction, cd, neutropenia, bacteremia, remission, relapsed, hospitalization, ms, consolidation, hidac, complete, febrile, relapse</td>
</tr>
</tbody>
</table>

Table 2. Diagnosis code-based representation of terms

<table>
<thead>
<tr>
<th>Term</th>
<th>aml sjogren shock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated Diagnosis codes (top 5)</td>
<td>205.00 : Acute myeloid leukemia without mention of remission 205.0 : Myeloid leukemia, acute 996.85 : Complications of bone marrow transplant 205.01 : Acute myeloid leukemia in remission 009.1 : Colitis, enteritis, and gastroenteritis of presumed infectious origin</td>
</tr>
</tbody>
</table>

Discussion

We presented a novel probabilistic model, HILDA, for jointly modeling clinical texts and diagnosis codes. This model generates term based representations of diagnosis codes and, in particular, develops interpretable representations for non-billing ICD-9 codes that never actually appear in the training data; i.e., even though these codes had no associated documents, the model is able to assign them interpretable distributions over terms (Table 1). Note that a query for “shock” returns comorbidities of shock such as respiratory failure and liver injury. Lastly, this model can also be used bi-directionally, identifying the most relevant diagnosis codes for a given query term. (Table 2) Thus, the model can be useful as a component in systems that perform tasks such as hypothesis generation, search, document and patient summarization, as well as automated clinical document labeling.

References


Acknowledgments We acknowledge National Science Foundation award #1344668
Information Blocking: How common is it and what policy strategies can combat it?

Eric Pfeifer, BA, School of Information and School of Public Health, University of Michigan, Ann Arbor, MI

Julia Adler-Milstein, PhD, School of Information and School of Public Health, University of Michigan, Ann Arbor, MI

Introduction: Congress has expressed concern about “information blocking” - the practice of knowingly interfering with the electronic exchange of patient health information – as a serious impediment to realizing value from EHRs and related HITECH investments. (1) However, there is little empirical evidence on who is engaging in information blocking, the specific forms it takes, and what policy solutions might effectively counter it. We therefore undertook a national survey of leaders of Health Information Exchange efforts (HIEs) to collect the first systematic data on information blocking.

Methods: Between October and December 2015, we conducted a nationwide, close ended survey of those currently leading state and community HIE efforts, typically CEOs, who would be most likely to observe information blocking if it were occurring. Respondents were those who participate in our annual census of HIE efforts. (2) The survey asked about several dimensions of information blocking (using the ONC definition of information blocking) separately for the two stakeholders currently suspected of information blocking: EHR vendors and hospitals/health systems. We created descriptive statistics to capture (1) the frequency of engaging in information blocking, (2) the specific forms it takes and degree of associated harm in terms of interfering with needed information sharing, and (3) the viability of policy solutions to counter it. The survey was initially piloted with three HIE leaders and was refined based on the feedback from these individuals.

Results: We received 60 responses (57% response rate). 50% of respondents reported that EHR vendors routinely engage in information blocking, and 25% of respondents reported that hospitals/health systems routinely do so. (Figure 1) The most common form of information blocking among vendors was deploying products with limited interoperability (49% indicated that vendors did so routinely, with an additional 31% reporting sometimes, Table 1) and this was also perceived as the most harmful form of information blocking (90% indicated very or moderately harmful). For hospitals/health systems, the most common form of information blocking was coercing providers to adopt a particular EHR or HIE technology (28% routinely and 24% sometimes, Table 1), but the most harmful form was using HIPAA as a barrier to HIE when it is not (83%). The forms that information blocking manifests itself from hospitals/health systems appears to be influenced by a fear of losing a competitive edge as a result of data sharing. To address vendor information blocking, the top policy action was tougher demonstrations of interoperability for certification (90% deemed very or moderately effective) and for hospital/health systems it was stronger CMS incentives to engage in HIE (92% deemed very or moderately effective).

Discussion: Based on the experiences of those leading HIE efforts, information blocking is widespread, though more common among vendors than providers. This is a key limitation of our study, we are only measuring the perception of whether information blocking is occurring. When information blocking occurs, it is widely perceived as harmful to needed information sharing. Viable policy remedies exist and the most popular would rely on existing mechanisms: EHR certification requirements and CMS incentives. Though there has been substantial discussion about information blocking since Congress raised it as an issue, it has been difficult to take action given that the evidence is anecdotal. Our study offers the first national data on the extent of information blocking as well as more detailed information on its forms, degree of harm, and viability of policy strategies to counter it. The federal government is in a position to take action and could do so without new legislation.
Figure 1. Frequency of Information Blocking

Table 2. Specific Forms of Information Blocking (Percent of Respondents)

<table>
<thead>
<tr>
<th>EHR VENDORS</th>
<th>Often/Routinely</th>
<th>Sometimes</th>
<th>Rarely/Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deployment of products with limited interoperability</td>
<td>49%</td>
<td>31%</td>
<td>20%</td>
</tr>
<tr>
<td>High fees for HIE unrelated to cost</td>
<td>47%</td>
<td>40%</td>
<td>13%</td>
</tr>
<tr>
<td>Making third party access to standardized data difficult</td>
<td>42%</td>
<td>41%</td>
<td>17%</td>
</tr>
<tr>
<td>Refusing to support HIE with specific vendors or HIEs</td>
<td>31%</td>
<td>37%</td>
<td>32%</td>
</tr>
<tr>
<td>Making data export difficult</td>
<td>28%</td>
<td>40%</td>
<td>32%</td>
</tr>
<tr>
<td>Changing HIE contract terms post-implementation</td>
<td>19%</td>
<td>21%</td>
<td>60%</td>
</tr>
<tr>
<td>Unfavorable contract terms for HIE</td>
<td>17%</td>
<td>35%</td>
<td>48%</td>
</tr>
<tr>
<td>Gag clauses on providers speaking out about Information Blocking practices</td>
<td>12%</td>
<td>18%</td>
<td>70%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HOSPITALS AND HEALTH SYSTEMS</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coercing providers to adopt particular EHR or HIE technology</td>
<td>28%</td>
<td>24%</td>
<td>48%</td>
</tr>
<tr>
<td>Controlling patient flow by selectively sharing patient information</td>
<td>22%</td>
<td>24%</td>
<td>54%</td>
</tr>
<tr>
<td>Using HIPAA as a barrier to PHI sharing when it is not</td>
<td>15%</td>
<td>35%</td>
<td>50%</td>
</tr>
</tbody>
</table>

References

Patient-provided Data Improves Race and Ethnicity Data Quality in Electronic Health Records

Fernanda Polubriaginof MD, MA1, Hojjat Salmasian MD, MPH, PhD2, Andrea Wells Shapiro MHA2, Jennifer Prey MPhil, MS1, George Hripcsak MD, MS1, Adler Perotte MD1, Nicholas P. Tatonetti PhD1, David K. Vawdrey PhD2,1

1Department of Biomedical Informatics at Columbia University, New York, NY;
2NewYork-Presbyterian Hospital, New York, NY

Introduction

Race and ethnicity are commonly used for estimating disease risk and for assessing health disparities, and these characteristics are frequently reported in observational studies that rely on Electronic Health Record (EHR) data. In the United States, the Meaningful Use financial incentive program includes the requirement that EHRs collect patients’ race and ethnicity in a structured fashion.1 Previous research on the quality of race and ethnicity data recorded in EHRs was conducted with small groups of selected patients.2,3 We undertook a study to evaluate data quality of race and ethnicity in EHRs, and to understand the impact of allowing patients to directly provide or curate their information.

Methods

We conducted a retrospective analysis of race and ethnicity data recorded for patients that had at least one visit at NewYork-Presbyterian Hospital/Columbia University Medical Center from January 1, 2014 to December 31, 2015. The institution collected race and ethnicity data using a two-question format, with one field recording the patient’s race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Unknown, Other, or Declined to Answer), and the second field capturing the patient’s ethnicity (Hispanic or Latino, Not Hispanic or Latino, Declined to Answer, Unknown). We calculated descriptive statistics on the frequency of race-and-ethnicity pairs. Patients classified as “Unknown”, “Other” or “Declined to Answer” were considered to have clinically uninformative data; we combined these categories into a larger category designated as “Unidentified”. Completeness was assessed based on the percentage of pairs with race and ethnicity not as “Unidentified” in the database. Furthermore, we analyzed the changes for race and ethnicity reported for the same patient over multiple visits, using HL7 logs from the EHR.

We assumed patient-reported data to be the reference standard for race and ethnicity data collection. In order to assess differences between patient-reported race and ethnicity information and the EHR data, we analyzed two sources of patient-reported data: 1) the HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems) Survey, which collected demographic information from patients in the form of a document sent via U.S. Mail after discharge, and 2) data from a randomized controlled trial where patients admitted through the hospital’s emergency department reviewed and corrected their race and ethnicity information (obtained via the EHR) using a custom patient portal application on a tablet computer.

Results

We found 352,486 unique patients who had at least one visit during our two-year study period. As shown in Table 1, 58.9% of patients did not have race or ethnicity identified in the EHR. There were 254,458 modifications made to race or ethnicity fields in the EHR for 153,535 unique patients. Modifications in the race or ethnicity could improve completeness (i.e., a change from ‘uninformative’ to a specific race or ethnicity category) but this was not always the case (Table 2). In fact, 30.5% of the modifications to the race field were changing “Other” to “Unknown”. Similarly, 29.7% of changes made to ethnicity converted “Hispanic” to “Unknown”.

During the study period, 28,658 unique patients responded to the HCAHPS survey, and 19,414 (67.74%) provided race and/or ethnicity information that was discordant with data recorded in the EHR. In the randomized trial that assessed patient-reported demographic data entered at the time of hospital admission, 65 patients were recruited. Of those, 35 (53.85%) made changes to their race and/or ethnicity (30 patients edited both, four patients edited ethnicity only, and one patient edited race only). Analysis of all of the “Uninformative” categories for race and ethnicity demonstrated that the majority of patients were willing to provide their information. Among the randomized trial study patients, 32 had “Uninformative” race information and 42 had “Uninformative” ethnicity documented in the
During, or after their healthcare encounters, could markedly improve data quality. Populations. Our findings suggest that patient-facing tools that allow patients to provide this information before, question format (i.e., collecting race and ethnicity data collection as separate fields) that is now widely used, as well as the clarity of the difference between “race” and “ethnicity” for patients. Interestingly, our Hispanic patients did not seem to consider themselves to belong to any of the OMB-defined race categories as the majority identified their race as “Other” and often entered “Hispanic,” the majority did not consider their race to be available in the list of options; 10 patients recorded or retained their race as “Other,” and two as “Unknown.”

Table 1. Frequency of race and ethnicity pairs in an electronic health record.

<table>
<thead>
<tr>
<th>Race</th>
<th>Ethnicity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>Uninformative</td>
<td>0.97%</td>
</tr>
<tr>
<td>Asian</td>
<td>Not Hispanic</td>
<td>0.96%</td>
</tr>
<tr>
<td>Asian</td>
<td>Hispanic</td>
<td>0.14%</td>
</tr>
<tr>
<td>Black</td>
<td>Not Hispanic</td>
<td>6.54%</td>
</tr>
<tr>
<td>Black</td>
<td>Hispanic</td>
<td>2.88%</td>
</tr>
<tr>
<td>Black</td>
<td>Uninformative</td>
<td>2.26%</td>
</tr>
<tr>
<td>Uninformative</td>
<td>Uninformative</td>
<td>27.93%</td>
</tr>
<tr>
<td>Uninformative</td>
<td>Hispanic</td>
<td>18.19%</td>
</tr>
<tr>
<td>Uninformative</td>
<td>Not Hispanic</td>
<td>4.52%</td>
</tr>
<tr>
<td>White</td>
<td>Not Hispanic</td>
<td>17.35%</td>
</tr>
<tr>
<td>White</td>
<td>Hispanic</td>
<td>12.79%</td>
</tr>
<tr>
<td>White</td>
<td>Uninformative</td>
<td>4.96%</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Table 2. Most frequent modifications to race/ethnicity in an electronic health record.

<table>
<thead>
<tr>
<th>Previous Race</th>
<th>Previous Ethnicity</th>
<th>New Race</th>
<th>New Ethnicity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>Hispanic</td>
<td>Unknown</td>
<td>Hispanic</td>
<td>10.29%</td>
</tr>
<tr>
<td>Other</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Hispanic</td>
<td>8.06%</td>
</tr>
<tr>
<td>Unknown</td>
<td>Unknown</td>
<td>White</td>
<td>Not</td>
<td>5.12%</td>
</tr>
<tr>
<td>White</td>
<td>Hispanic</td>
<td>Unknown</td>
<td>Unknown</td>
<td>4.40%</td>
</tr>
<tr>
<td>Unknown</td>
<td>Hispanic</td>
<td>Unknown</td>
<td>Unknown</td>
<td>3.29%</td>
</tr>
<tr>
<td>Unknown</td>
<td>Unknown</td>
<td>White</td>
<td>Unknown</td>
<td>2.94%</td>
</tr>
<tr>
<td>Other</td>
<td>Hispanic</td>
<td>Unknown</td>
<td>Unknown</td>
<td>2.89%</td>
</tr>
<tr>
<td>Other</td>
<td>Hispanic</td>
<td>White</td>
<td>Hispanic</td>
<td>2.70%</td>
</tr>
<tr>
<td>Other</td>
<td>Unknown</td>
<td>Other</td>
<td>Hispanic</td>
<td>2.43%</td>
</tr>
<tr>
<td>White</td>
<td>Hispanic</td>
<td>Other</td>
<td>Hispanic</td>
<td>2.28%</td>
</tr>
</tbody>
</table>

Discussion

The U.S. Meaningful Use of EHRs financial incentive program requires participating healthcare providers to record patients’ race and ethnicity in a structured fashion. Categories for collecting race and ethnicity are based on current standards published by the Office of Management and Budget (OMB) in 1997. When analyzing race and ethnicity data in the EHR, we observed that a significant percentage of patients had a documented race and ethnicity that was not clinically informative. Analysis of changes in race and ethnicity data over time showed that the most frequent changes were from uninformative concepts to different uninformative concepts (e.g., from ‘Other’ to ‘Unknown’). We also observed that changes were made not only when uninformative concepts were detected, but also from more informative to less informative concepts (e.g., from ‘White Hispanic’ to unknown race, unknown ethnicity). This phenomenon speaks to the challenges faced by registration personnel, particularly in care delivery settings with racially and ethnically diverse populations.

One way to improve the quality and completeness of patient demographics in electronic health records is to allow patients to review and request updates to their own information. This could be accomplished in many ways, but one useful method may be through patient portals. When comparing ethnicity data before and after patient review in the Emergency Department, we observed that patients were willing to review their information and make changes when needed. The majority of the patients with uninformative race and ethnicity in the EHR changed these values to more meaningful concepts. Interestingly, our Hispanic patients did not seem to consider themselves to belong to any of the OMB-defined race categories as the majority identified their race as ‘Other’ and often entered “Hispanic”, “Latino” or their country of origin in a free-text field. This behavior raises questions about the efficacy of the two-question format (i.e., collecting race and ethnicity data collection as separate fields) that is now widely used, as well as the clarity of the difference between “race” and “ethnicity” for patients.

The appropriate collection of race and ethnicity information is key to recognizing disparities that affect minority populations. Further, this information can be used to perform disease risk assessment both for individuals and populations. Our findings suggest that patient-facing tools that allow patients to provide this information before, during, or after their healthcare encounters could markedly improve data quality.

References

Validating the Occupational Data for Health Model: An Analysis of Occupational Information in Reports, Standards, Surveys, and Measures

Sripriya Rajamani, MBBS, PhD, MPH1, 3, Elizabeth S. Chen, PhD2, Ranyah Aldekhyyel, MS3, 5, Yan Wang, PhD3, Genevieve B. Melton, MD, PhD3, 4
1Public Health Informatics Program, 3Inst for Health Informatics, 4Dept of Surgery, Univ of Minnesota, Minneapolis, MN; 2Ctr for Biomedical Informatics, Brown Univ, Providence, RI; 5Medical Education Dept, College of Medicine, King Saud Univ, Riyadh, SA

Introduction
Recognizing the importance of social factors in influencing health status and outcomes, recent reports have advocated for their incorporation in electronic health record (EHR) systems. The National Academy of Medicine (NAM; formerly Institute of Medicine) 2011 report1 titled “Incorporating Occupational Information in Electronic Health Records” highlighted the need for robust representation of occupational information in EHRs. Work initiated by the National Institute for Occupational Safety and Health (NIOSH), a leader in promotion of occupational health and strong supporter for incorporation of occupational information in EHRs, has resulted in an ODH (Occupational Data for Health) data model2. Subsequent statements by reputed sources3-5 as well as landmark reports by NAM in 2014 on social and behavioral domains and measures in EHRs6,7, further emphasize the importance of capturing occupation-related information. NIOSH has also supported demonstration projects around occupational data. The study objective was to collect occupational health items from resources ranging from recommendations, standards, public health reports, public health surveys, and research measures in order to validate the ODH model.

Methods
Five key source types (reports and recommendations, standards and specifications, public health reports, public health surveys, and research measures) comprising 20 resources were identified by review of relevant reports and literature searches (e.g., using “occupation” and “representation” as search terms) to create a master list of items (e.g., elements or questions/responses) on representation of occupational information. This list was used as a guideline to evaluate the ODH model by NIOSH which comprises of six categories: Occupational History, Usual Occupation and Industry, Employment Status, Work Schedule, Occupational Injury, and Occupational Exposure. The methodology consisted of four main components: (1) create a comprehensive list of items from representative sources; (2) develop guidelines to map items to the ODH model; (3) map items to ODH categories and elements; and (4) ascertain coverage and identify areas for enhancement. Guidelines for mapping were developed by study authors with expertise in biomedical standards, informatics, and clinical and public health practice. Two of the experts (SR and RA) mapped 10% of items calculating inter-rater reliability (Cohen’s kappa = 0.94 and proportion agreement = 0.99) and also mapped all items, which were then reviewed and analyzed by the study authors.

Results
Twenty resources (Table 1) including 247 items were identified for inclusion in the master list for representation of occupational information and 100% mapped to the ODH model. Occupational History was the most common category across the resources (65%), followed by Employment Status (55%), and Usual Occupation and Industry (45%). Certain elements contained clinical interpretations (e.g., Work-relatedness) and some were derived data elements (e.g., duration of Occupation and Industry). The need for more granularity in the Occupational Injury and Occupational Exposure categories was identified. Some of the survey questions captured temporality for recent past (e.g., last week) and compensation for work (e.g., wages). The ODH model includes Employer Location that may be different from actual worksite for exposure data. Value sets for Employment Status were varied across the resources.

Discussion
Our study findings illustrate the robustness of the ODH model for representation of occupational information that may ultimately contribute to supporting clinical care decisions. The results also highlight the multitude of associated concepts and complexity of their representation. The ODH model is in draft stages (under review) and likely to have expanded elements (e.g., for Occupational Exposure) when released. Definitions of the occupational categories and elements, codes for their representation, and guidance for certain categories (e.g., Employment Status) are critical for uniformity in use. With the growing adoption of EHRs and the possibility of future EHR certification criteria for occupational information, it is essential for reputed entities like NIOSH to continue their advocacy.
Table 1: Mapping of Resource Items to the NIOSH Occupational Data for Health (ODH) Model

<table>
<thead>
<tr>
<th>Categories and Elements from the ODH Model</th>
<th>Reports and Recommendations</th>
<th>Standards and Specifications</th>
<th>Public Health Reports</th>
<th>Public Health Surveys</th>
<th>Research Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupational History</td>
<td>2011 NAM</td>
<td>2014 NAM #1</td>
<td>IHE PCC TF</td>
<td>PHIP-ODIF</td>
<td>HL7 FHIR</td>
</tr>
<tr>
<td></td>
<td>2014 NAM #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2014 NAM</td>
<td></td>
<td>Death Certificate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry Description</td>
<td></td>
<td></td>
<td>PHFP ODIF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job Duties</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current Occupation Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>End Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hours Worked per Week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days Worked per Week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employer Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employer Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual Occupation and Industry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration in Years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>End Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work Schedule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupational Injury</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupational Exposure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NAM - National Academy of Medicine (formerly Institute of Medicine); IHE PCC TF - Integrating the Healthcare Enterprise Patient Care Coordination Technical Framework; PHIP - Public Health Functional Profile; ODF - Occupational Data, Injuries and Fatalities; NHANES - National Health and Nutrition Examination Survey; BRFSS - Behavioral Risk Factor Surveillance System; SOII - Survey of Occupational Injuries and Illnesses; PhenX - Consensus Measures for Phenotypes and eXposures; CDE - Common Data Elements

Acknowledgements

The authors would like to thank Dr. Genevieve Luensman (NIOSH) for providing context on the NIOSH ODH model. This work was supported in part by National Library of Medicine grant R01LM011364.

References

Information Needs of Physicians, Care Coordinators, and Families to Support Care Coordination of Children with Special Health Care Needs (CSHCN)

Pallavi Ranade-Kharkar, PhDC,1,2 Charlene Weir, PhD, RN2,3, Chuck Norlin, MD2,4, Sarah A. Collins, RN, PhD5, Lou Ann Scarton, PhDC, Gina B. Baker, RN, MSN1, Damian Borbolla, MD, MS2, Vanina Taliercio, MD, MS2, Guillerme Del Fiol, MD, PhD2
1Intermountain Healthcare, Murray, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 3VA Medical Center, Salt Lake City, UT; 4Department of Pediatrics, University of Utah Health Sciences Center, Salt Lake City, UT; 5Partners HealthCare, Brigham and Women’s Hospital and Harvard Medical School, MA.

Introduction
Children with special health care needs (CSHCN) are “those who have or are at increased risk for a chronic physical, developmental, behavioral, or emotional condition and who also require health and related services of a type or amount beyond that required by children generally.”1 CSHCN comprise an estimated 15.6% (11 million) of the pediatric population in the US, and their healthcare costs 3 times that of other children1. Caring for CSHCN is highly complex and often involves multiple medical and non-medical providers working independently and using disparate information systems. As a result, finding the information needed to support the care of CSHCN can be daunting. Timely access to information can support care coordination for CSHCN, avoiding care delays, duplicated efforts, diminished quality and frustration, and bring substantial value. Previous studies on care coordination for complex patients identified care process activities. In this study we aimed to elicit the kinds of information needs and associated goals that are raised by physicians, care coordinators, and family members while coordinating the care for CSHCN. Ultimately, our findings can be used to inform the design of informatics tools to support and improve care coordination for CSHCN.

Methods
We conducted interviews according to the Critical Decision Method (CDM); triangulating information from physicians, care coordinators, and family members (University of Utah IRB# 00075524). Interviewees were recruited from practices in a pediatric Patient-Centered Medical Home Demonstration project in Utah and represented different healthcare institutions. We interviewed 8 physicians, 7 families and 4 care coordinators; total interview time was over 17 hours. Following the CDM technique, interviewees were asked to recall a recent episode of care for CSHCN that they perceived to be challenging from the information-seeking perspective. The interview script prompted for goals (why), content (what and from whom), mode (how), barriers, and strategies used to meet the information needs. We systematically analyzed the interview transcripts using a content analysis method.3 Interviews were first unitized based on closely related sentences forming a meaning unit. From each unit, two researchers independently extracted one or more information needs in the form of questions, based on a code book. These questions were then converted to generic questions (e.g.,“What are the child's seizure patterns?”) and iteratively classified into information need goals. Throughout the process, disagreements were resolved through consensus among the researchers.

Results
On average, each interview identified 80 information needs. Our analysis categorized them into six information need goals: 1) situation understanding, 2) care networking, 3) planning, 4) tracking/monitoring, 5) navigating the health care system, and 6) learning (Table 1). Physicians and care coordinators described the need to develop situation understanding about the patient through accessing extensive healthcare records from various settings. For families, this goal related to understanding the assessments and goals of clinicians throughout the care process. Interviewees described care networks for CSHCN that go much beyond the healthcare system and include schools, community services, extended families and friends. Physicians coordinated care within the primary, specialty and the extended care network. Families reported frustrations due to inadequate feedback from their child’s care network. Physicians described planning for potential life-threatening emergencies. Families spoke of vigilantly planning episodes of care with the goals of optimizing resource utilization and trying to maintain a sense of normalcy in their child’s life. Physicians monitored test results and trends of assessments, care coordinators tracked episodes of care and completion of tasks. Families spent significant time and effort tracking symptoms, diet and response to medications. Parents spoke of having to navigate complex and time-sensitive processes for procuring needed services and resources with no single source of information. Care coordinators greatly assisted parents by taking care of paperwork-intensive processes to ensure coverage of care and regulatory requirements. Families had a significant
need for learning about symptoms, side-effects, and alternative treatments. Interviewees reported that meeting these information needs typically required considerable effort and involved various individuals and multiple information systems. Quite often, families had the burden of serving as the coordinators of their child’s health care, seeking and conveying information among members of their care network.

Discussion
To our knowledge, this is the first study to provide an in-depth assessment of information needs and associated goals in the care of CSHCN. We found that caring for these children generates a large number of information needs that require substantial effort from families, physicians, and care coordinators. These information needs present significant opportunities to improve care through multifaceted informatics solutions (e.g. a HIPAA-compliant social network application for care networking), integrated with multiple systems via health information exchange (HIE).

Table 1. Information need goals, definitions and corresponding examples of quotes from interviews.

<table>
<thead>
<tr>
<th>Goals</th>
<th>Definitions</th>
<th>Example Quotes From Interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation understanding</td>
<td>Creating a mental model of the patient by integrating pieces of information about the patient, environment, history, and preferences.</td>
<td>“So the challenges on the first visit, I’d say there were a few. First of all, no prior records available made it kind of tough. So not really having a good grasp of how was this kid growing. When the parents came, they had requested them ahead of time. They never arrived. We didn’t have them and we couldn’t get them.”</td>
</tr>
<tr>
<td>Physician quote</td>
<td>Building a patient’s care team, knowing team member identities and roles, and sharing pertinent information.</td>
<td>“We need to have a clearer relationship with the neurology and communicate more clearly with neurology so that they know what we're doing and we know what they're doing, so we can work together because we can't rely on a parent to pass on that information back and forth.”</td>
</tr>
<tr>
<td>Planning</td>
<td>A process that starts with choosing healthcare goals, followed by evaluating alternate routes, and finally developing a specific plan.</td>
<td>“We had planned (her back-to-back surgeries) about a year ahead of time. We looked at her school schedule and … tried to coordinate that with when she would be off track, so she would miss the least amount of school” “In the six months preceding the surgery, we worked really closely with (the dysphagia clinic) to bump up her calories and her feeds to try and bulk her up ahead of the surgery.”</td>
</tr>
<tr>
<td>Care networking</td>
<td>Following adherence to and execution of the treatment plan, the patient’s progression towards care goals, and effect of treatment on outcomes.</td>
<td>“How many seizures she had had in the past two or three months? How many is typical for her to have and like the situations around the seizures like were they explained? … I track the date, the time, how long the seizure lasted, if there was a trigger that I thought caused the seizure and then if I needed to use her rescue meds.”</td>
</tr>
<tr>
<td>Care coordinator quote</td>
<td>Understanding and executing logistical and process tasks, typically related to navigating the health care system.</td>
<td>“This client does wear diapers. So being in our program, &lt;name removed&gt; will make a prescription for the diapers, and we’ll order those because the funding – I mean they’ll get paid through the home program Medicaid.”</td>
</tr>
<tr>
<td>Learning</td>
<td>Seeking information about or getting educated on the various aspects and ways of caring for CSHCN.</td>
<td>“You know, I researched most things on my own because I just realize that doctors are only humans, and I know my daughter best. I know what she responds, what she likes, how she improves with her health. I know that she does much better without medication than with medication. So, I have chosen very non-traditional ways.”</td>
</tr>
</tbody>
</table>

References
Design and Implementation of an Ancillary Genomics System for the Return of Pharmacogenetic Results

Luke V. Rasmussen¹, Maureen E. Smith¹, Federico Almaraz², Stephen D. Persell¹, Laura J. Rasmussen-Torvik¹, Jennifer A. Pacheco¹, Carl Christensen¹,², Timothy M. Herr¹, Firas H. Wehbe¹, Justin B. Starren¹

¹Northwestern University Feinberg School of Medicine, Chicago, IL; ²Northwestern Memorial HealthCare, Chicago, IL

Introduction

Genomic laboratory tests have been an established tool for the diagnosis and treatment of certain conditions. With more recent advances in sequencing technology and in our understanding of the human genome, these tests have offered new options for patient care. Historically, the results of these tests have been returned as narrative text reports that are then incorporated into the patient’s medical record. They may be stored in locations that do not facilitate future retrieval, and in formats that are not computable for use in clinical decision support (CDS).

Multiple approaches have been put forth, each with strengths and limitations¹, to enable the flow of genomic information from the lab to the clinician. To date, there exist several examples of genomic medicine programs that have integrated genomic lab results with the electronic health record (EHR)²⁻⁵; however, these do not represent all possible implementation paths. Following the design principle that genetic and genomic data require special handling (similar to picture archiving and communication systems [PACS] for radiology images) and are best suited for storage in an ancillary system rather than the core EHR, we designed and implemented the Ancillary Genomics System (AGS) as part of the electronic Medical Records and Genomics Pharmacogenomics (eMERGE PGx) project⁶.

Methods

As part of eMERGE PGx, Northwestern University selected three drug-gene variant interactions (DGIs) of interest for which samples would be analyzed and returned in a primary care setting (clopidogrel: CYP2C19; warfarin: CYP2C9 and VKORC1; simvastatin: SLCOB1). Two external CLIA laboratories processed samples from 750 consented subjects. The Northwestern IRB reviewed and approved the study.

Although focused on a PGx use case, we designed the AGS to receive and process lab results for different levels of genotyping (e.g. from SNP panels to sequencing). Furthermore, the AGS was designed to house the original data from the CLIA lab, to process it using a set of approved rules, and to create a derived interpretation (or a “computed observation”) that would then be transmitted to the EHR as a discrete result. The intent of this design is to streamline the development of CDS alerts and rules. As knowledge or clinical recommendations change over time, the system design will allow existing data to be reprocessed, and updated results transmitted to the EHR.

Figure. The Ancillary Genomics System, including sub-components and interactions with external systems.

Results

An overview of the AGS design is shown in the Figure. It consists of various sub-components interacting with a central data store. The Data Import module is a collection of data adapters developed for the specific format provided by each CLIA lab for this study. Each adapter takes an import file and extracts the discrete results for storage in a relational database. The Data Analysis component includes a separate interpretation engine for each
Each engine first normalizes the results depending on the original format of the lab result (e.g., star variant, SNP genotype). For example, a homozygous star variant represented as just “*2” would be normalized to [‘*2, *2’]. This allows for more streamlined and consistent processing of results. The analysis module then uses these normalized results along with a set of lookup tables to create computed observations representing the corresponding PGx phenotype (e.g., “clopidogrel poor metabolizer”). In addition to a discrete representation of the phenotype, the system generates an interpretation report describing the corresponding finding and its implications.

Additionally, we developed a web client to allow study staff to review the computed observations and associated interpretation, and to flag unusual results not conforming to the standard interpretation. The system was initially developed to allow two phases of approval – the first to verify and approve the result, and the second to release approved results to the EHR. Study staff can control when and how results were returned, with an option to automatically release to the EHR upon approval. Upon release, the Results Export component transmitted the computed observation and report to the EHR. Within our configuration, each observation was returned as a lab result using a standard HL7 v2 message, although any HL7 v2 message could be generated.

Following withdrawals during the course of the study, in total 746 patients had each of their 3 DGI results returned to the EHR (2238 results in total).

Discussion

Although other approaches to integrating genomic results within the EHR have been successful, we highlight here the following benefits of the AGS solution presented above: (1) A dedicated AGS for storing lab results reduced the complexity of reviewing genomic results for most clinicians. This AGS design allowed us to centralize review on the computed observation, and then link to supporting genomic results. (2) The upfront use of computed observations reduced the burden on the EHR CDS engine by not requiring a large number of decision tree paths to be explored each time a patient’s record was viewed. While only 746 patients were included in this study, we anticipate this approach to scale in the future to incorporate more patients, more DGIs, and more complex results. (3) The use of computed observations allowed the CDS rules to be more streamlined (requiring fewer Boolean combinations and decision paths). This reduced burden on clinical analysts during the CDS setup, and required no intervention from clinical analysts during the study as new variant combinations were presented. (4) The modular and independent computation of observations easily permits re-interpretation and delivery of new result “interpretation” in the event knowledge about the variant changes.

Conclusion

Following the principles applied to the specialized handling of radiology images in PACS, we have designed and successfully implemented an Ancillary Genomics System at our institution for returning PGx results to patients in the EHR. The use of an AGS may benefit other institutions wishing to manage discrete genomic results returned from external laboratories.

References

Technologies to Support Care Coordination for Medication Safety

Alissa L. Russ, PhD1-4, Cherie L. Luckhurst, PhD,1,2 Rachel A. Dismore, MA,1,2 Karen J. Arthur, PharmD,2 Amanda P. Ifeachor, PharmD,2 Peter A. Glassman, MBBS, MSc,5 Michael Weiner, MD, MPH1-3
1Center for Health Information and Communication, Department of Veterans Affairs (VA), Health Services Research and Development Service, 2Richard L. Roudebush VA Medical Center, and 3Regenstrief Institute, Inc., Indianapolis, IN; 4College of Pharmacy, Purdue University, West Lafayette, IN; 5VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA.

Introduction
Care coordination, the intentional organization of patient care activities between two or more individuals,1 is essential for medication safety. Limited research, however, has been conducted on care coordination as it applies to medication safety incidents. A recent, systematic review indicates a need for research on these topics.2 Our objective was to identify factors that make it difficult for healthcare professionals to coordinate care when resolving medication safety incidents. We also conducted a human factors analysis to identify approaches that could ease or improve coordination around such incidents.

Methods
As part of a larger study, physicians and pharmacists from a major Veterans Affairs Medical Center submitted cases, via a study tool, where they identified and addressed a medication safety incident. We completed cognitive task analysis interviews for 60 cases total across 3 categories: 1) adverse drug reaction; 2) drug-drug interaction; and 3) drug-disease interaction involving renal function. Cases were eligible for inclusion in the analysis if participants coordinated with any other individuals to resolve the incident, yielding 47 eligible cases. We stratified eligible cases by the three categories above, and by participants’ professional role (i.e., physician or pharmacist), to ensure even sampling. Since qualitative data saturation is typically achieved within a smaller sample size, 24 cases were randomly selected from the eligible cases for analysis, with plans to analyze additional cases, if needed, to reach adequate saturation. We performed an inductive, qualitative analysis without predetermined themes,3, 4 to identify factors that hindered healthcare professionals’ coordination of care. The analysis team (human factors expert, pharmacist, sociologist, and research assistant) identified emergent themes and discussed any coding discrepancies until reaching consensus.5

Results
We analyzed 24 cases (from 12 physicians, 12 pharmacists) with 8 cases from each category, above. No new themes emerged after the first 12 cases, indicating adequate data saturation. Overarching themes are outlined in Table 1. Results from the first theme suggest that healthcare systems should facilitate real-time communication for some types of coordination. Other results suggest five potential technology-related solutions to enhance coordination: 1) technologies to facilitate more directed, medication-related communication among healthcare professionals; 2) prescribing guidelines and drug compendia that are integrated into computerized provider order entry (CPOE) systems; 3) a semi-automated CPOE design that allows healthcare professionals to enter safety criteria for individual medication orders, so that if these criteria are met (e.g., a lab test exceeding a designated threshold), the prescriber and pharmacy would be automatically notified and, if pre-requested, the medication would be automatically discontinued; 4) documentation mechanisms, tied to specific medication orders, that state the medication’s indication and can be easily viewed by the patient’s other providers; 5) documentation mechanisms, tied to specific medication orders, that explain safety-related changes to individual orders, are easily viewable by the patient’s other healthcare providers, and that persist in the EHR long-term.

Conclusions
To our knowledge, this is the first study to systematically analyze care coordination activities that occurred in actual medication safety incidents. Results reveal factors that impede healthcare professionals’ ability to coordinate care for medications. Recommendations from this research are hypotheses that would need further evaluation. Results are expected to inform healthcare system design changes, some of which may be technology-based, to aid care coordination activities; findings may ultimately be used to enhance medication safety for patients.
Table 1. Factors that hindered healthcare professionals’ coordination of care when resolving medication safety incidents.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Definition</th>
<th>Example(s)</th>
</tr>
</thead>
</table>
| 1. Sole reliance on IT/challenges with IT | Challenges with design and use of information technology, including placing trust in electronic tools and using them as the sole means of communication. | “…it’s easy [for me] to type a little addendum to the [EHR] note. If there were a better way, maybe I could just go up to her directly maybe and it would’ve shaved two days off of resolving this.”

“I was having trouble sorting through the chart [EHR] and figuring out why it [omeprazole] had been stopped, and I couldn't figure out what they felt the drug interaction was.” |

| 2. Breakdowns | Delays or errors that occurred, including errors caused by misinformation, breakdowns related to handoffs between healthcare professionals, and activities that were planned but not completed. | “…that little delay [caused] like [a] week delay…where we were kind of going back and forth and deciding what to do for him.” |

| 3. Roles and Ownership | Problems or confusion around responsibilities for patient care activities. Examples include uncertainty, ambiguity, or confusion about who is responsible for making decisions or carrying out actions. This also includes challenges related to a healthcare professional’s scope of practice or authority. | “Why the heck did the primary care physician not make an executive decision? It’s her patient, too…. I was really frustrated that the primary care physician hadn’t just stopped the doxazosin.” |

| 4. Complexity of Coordination | Factors that increased the complexity of managing care for the patient. | “[The patient had already been discharged from the hospital when we discovered the medication issue], so we called the patient’s son, and the patient, and reviewed his discharge instructions because he was [already] gone!” |

References

Abstract: Mobile Apps for Vulnerable Populations Study
Urmimala Sarkar, MD, MPH1, Gato Gourley, MSc1, Courtney Lyles, PhD1, Lina Tieu, MPH1, Cassidy Clarity, BA1, Lisa Newmark, BA2, Karandeep Singh, MD, MMSc3, David W. Bates, MD, MSc4, 5, 6

1University of California San Francisco, Center for Vulnerable Populations at San Francisco General Hospital, United States of America
2Information Systems, Partners HealthCare System, Wellesley, MA, United States of America
3Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, United States of America
4Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, United States of America
5Harvard Medical School, Boston, MA, United States of America
6Department of Health Policy and Management, Harvard T. H. Chan School of Public Health, Boston, MA, United States of America

Introduction: Mobile applications or ‘apps’ intended to help people manage their health and chronic conditions are widespread and gaining in popularity. (1) However, little is known about their acceptability and usability for low-income, racially/ethnically diverse populations who experience a disproportionate burden of chronic disease and its complications. (2, 3) The objective of this study was to investigate the usability of existing mobile health applications (“apps”) for diabetes, depression, and caregiving, in order to facilitate development and tailoring of patient-facing apps for diverse populations.

Methods: Usability testing was conducted with participants only on condition concordant apps, with caregivers (n=9) tested on caregiver apps, participants with depression (n=10) tested on depression apps, and participants with type 2 diabetes (n=10) tested on diabetes apps. There were a total of 11 of the most popular health apps (4 diabetes apps, 4 depression apps, and 3 caregiver apps) on both iPad and Android tablets. Apps were selected by a panel of experts as the most optimal for supporting self-management and were selected based on high ratings in the Android and Apple app stores. Participants were given condition-specific tasks, such as entering a blood glucose value into a diabetes app. The task completion coding scheme was developed a priori using adapted usability metrics from prior studies. (4, 5) We identified the proportion of tasks that were completed independently; and the degree of completion, categorized as: a) successful/straight-forward, b) successful/prolonged, c) partial, unsuccessful/prolonged, and d) gave up. (4, 5)

Results: Participants were recruited from an urban safety net hospital. We did not ask participants about income level, but it is understood that patients at this urban safety net hospital are low income. (6, 7) The participants were diverse: 58% African American, 27% White, 8% Asian, 8% Latino. The majority of participants were over 50 years old, with over half 56 and older. The majority of participants had limited health literacy and over half had at least some interest in using the internet to manage their health. In addition the vast majority had experience using a computer. Participants completed 79/185 (43%) of tasks across 11 apps without assistance. Participants performed better on data entry tasks compared to data retrieval tasks (Figure 1). Three themes emerged from participant comments: lack of confidence with technology, frustration with design features and navigation, and interest in having technology to support their self-management (Table 1). Specific design features that participants found frustrating were button placement and data entry locations as well as difficulty of navigating through multiple screens.

Discussion: App developers should employ participatory design strategies in order to have an impact on chronic conditions like diabetes and depression which disproportionately affect vulnerable populations. While patients express interest in using technologies for self-management, current tools are not consistently usable for diverse patients.
Table 1

<table>
<thead>
<tr>
<th>Theme</th>
<th>Quotation (health condition, app name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frustration</td>
<td>&quot;What's interesting is, I'm a computer geek, and this is not friendly whatsoever.&quot; (Diabetes, Diabetes App)</td>
</tr>
<tr>
<td>Lack of Confidence</td>
<td>&quot;Like I said, a lot of people's not gonna, if you're not computer literate, you know, forget about it.&quot; (Diabetes, Diabetes Connect)</td>
</tr>
<tr>
<td>Interest in technology to support self-management</td>
<td>&quot;Yes I would. I sure would. How technology is, and how you can keep up with your sugar diabetes and how you can control with your sugar diabetes and how it can help you with your walking with your eating, with your health, and with everything about what's going on with your sugar diabetes type 2.&quot; (Diabetes, reflection on all apps)</td>
</tr>
</tbody>
</table>

References
Templates for Ordering Specialty Care Consultations Lack Minimalism and Error Prevention: Results of a Heuristic Evaluation

April W. Savoy, PhD1,2,3, Himalaya Patel, PhD1,2, Mindy Flanagan, PhD2, Alissa L. Russ, PhD1,3,4, Michael Weiner, MD, MPH1,2,3

1Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Richard L. Roudebush VA Medical Center, 2Indiana University, 3Regenstrief Institute, Inc., Indianapolis, IN; 4Purdue University, West Lafayette, IN

Introduction
Through electronic health record systems (EHRs), primary care providers (PCPs) frequently order consultations, also called consults, to refer patients to specialty clinics for evaluation or management. In the EHR used by the Veterans Health Administration (VHA), to facilitate consult ordering, specialty clinics have added templates for data entry in the consult order form. These templates extend the EHR’s user interface and organize the data saved in the order form. Some templates also communicate specialists’ requirements (e.g., pre-visit diagnostic testing) and may offer clinical guidance for both common and uncommon cases.

However, because each specialty clinic at each VHA facility maintains its own requirements and guidance—as may also occur among EHR users of multi-facility institutions outside the VA—consult templates lack standardization in structure and content. As a result, these templates vary in their efficiency and effectiveness, both within and among facilities. Incomplete or misdirected consult orders are frequently cancelled or discontinued without an appointment for the patient, creating delays in the delivery of care. The objective of this study is to identify usability problems in consult templates to guide design changes that will improve the efficiency and safety of care.

Methods
To identify the prevalence, variety, and potential severity of usability problems in VHA consult templates, a sample of templates was collected for heuristic evaluation, a method for assessing usability through expert evaluations of user interfaces (the templates, in this case). Templates were selected in a two-stage process. The first stage focused on templates identified directly by PCPs as either particularly easy or particularly difficult to use. The second stage focused on orders that were created but not executed (i.e., either cancelled or discontinued). The templates came from three geographically diverse VHA facilities. As part of a larger project to identify barriers and facilitators in the consultation process, the research team conducted 30-minute semi-structured interviews of primary care clinicians from December 2014 to February 2015 and from November 2015 to February 2016, at two of the facilities. From these interviews we marked explicit references to specialty clinics, yielding 11 corresponding templates. Data for the second stage covered consults from 2014 across all three facilities in six specialty clinics: cardiology, mental health, oncology, ophthalmology, orthopedics, and rheumatology. Because clinics often supplied multiple templates to accommodate specific clinical conditions, we excluded templates for routine procedures, inter-facility orders, out-of-network (non-VA) orders, and infrequent orders (less than 300). From those remaining, we selected the templates with the lowest completion rate (i.e., the highest combined cancellation and discontinuation rate). This stage yielded 18 templates with completion rates between 39% and 81%. Overall, of 28 templates selected, 26 were unique.

These templates were evaluated against 19 usability heuristics. Eighteen heuristics were from domain-independent lists published separately by Nielsen and Shneiderman. A nineteenth heuristic, meet users’ information needs, was created to address the difficulty of communication between clinicians. The evaluators were four researchers with backgrounds in social psychology, human factors engineering, or human–computer interaction. Evaluators viewed screenshots of each template and recorded each observed violation on a row of a spreadsheet. For each violation, evaluators recorded the most applicable usability heuristic, a description of the problem, applicable usability goal (efficiency, effectiveness, safety, learnability, memorability, and utility), and the estimated severity of the violation (integer, 0 to 4, with 4 indicating the most severe). Two evaluators reviewed each template for overlapping findings. Within a given template, findings overlapped if their descriptions identified the same basic problem—whether found multiple times by multiple evaluators, multiple times by one evaluator, or one time by multiple evaluators. Each group of overlapping findings was recoded into a single finding. For that recoded finding, the heuristic, dimension, and severity were determined by the group’s mode.
Results
According to the heuristic evaluation, the examined consult templates are needlessly complex. Recurring causes of heuristic violations included inconsistent organization, missing or misplaced instructions, instructions causing work outside the order form, and selections or actions with unclear outcomes.

The evaluators found 201 violations with an average of 7 violations per template. Of the 19 available heuristics, 14 were violated at least once. Five heuristics held most major and total violations (Table 1). An example of a major violation of recognition rather than recall was the requested reentry of data from elsewhere in the patient’s record. An example of a major violation of meet users’ information needs was a missing field for entering the reason for consultation. Of the six usability goals, five had potential negative impacts based on the heuristic violations. Efficiency and safety had the greatest potential impact based on major and total violations (Table 2).

Table 1. Examples of usability heuristics violated most often in the evaluated templates

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Major Violations</th>
<th>Total Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rather than recall</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>Meet users’ information needs</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Error prevention</td>
<td>9</td>
<td>33</td>
</tr>
<tr>
<td>Reduce short-term memory load</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Design dialogs to yield closure</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 2. Usability goals potentially impacted by violations

<table>
<thead>
<tr>
<th>Usability Goal</th>
<th>Major Violations</th>
<th>Total Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency</td>
<td>22</td>
<td>65</td>
</tr>
<tr>
<td>Safety</td>
<td>19</td>
<td>49</td>
</tr>
<tr>
<td>Learnability</td>
<td>3</td>
<td>44</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>Utility</td>
<td>7</td>
<td>20</td>
</tr>
</tbody>
</table>

Supplemental analysis indicated alignment among PCPs and evaluators regarding template quality. Evaluators’ comments echoed PCPs’ preference for templates with short length, clear guidelines, and linked order sets. A notable point of caution was the evaluators’ assumption of shared jargon between PCPs and specialty care providers. PCP comments indicated that unfamiliar terms were used in templates, which caused confusion.

Discussion
Although heuristic evaluations do not generate solutions to usability problems, they identify problems and describe the nature of violations to provide guidance for design changes. Our results indicate key principles for new template designs: concisely communicate specialty clinics’ requirements, limit context switches to other parts of the EHR with auto-population, and support PCPs’ need to communicate information outside the provided structured input fields. Concise communication of specialty requirements supports PCPs’ preferences for shorter templates and clear guidelines that would aid in increasing effectiveness of consult requests via templates. Limiting context switches would increase PCPs’ efficiency related to consult requests by eliminating excessive, redundant navigation to duplicate information in templates. Lastly, the inclusion of free-text fields enables PCPs to communicate additional information that is pertinent to the consult request, enhancing PCP-specialist communication and consult coordination. By addressing usability issues in the first step of the consult process, we expect a decrease in the number of consult requests that are discontinued or cancelled—problems that cause delays in care. Future work includes the evaluation of commercial templates to assess the applicability of the design principles to non-VA templates.

References
A Method for Enhancing the Portability of Electronic Phenotyping Algorithms: An eMERGE Pilot Study

Ning Shang, PhD, Chunhua Weng, PhD, George Hripcsak, MD, MS
Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract
This work proposes a method to facilitate eMERGE phenotypes’ implementation that is portable across heterogeneous data environments. Parameterized and modularized queries facilitate the portability by parameterizing terminology and data schema and separating them from the query logic. Common data elements are identified to satisfy core algorithm requirements. Altogether, this research provides insight for developing a common data model for electronic phenotyping.

Introduction
Electronic phenotyping uses data from Electronic Health Records (EHRs) to provide computational definitions of phenotypes, which can then be used for genome-wide association studies or clinical study cohort identification. Existing approaches to develop electronic phenotypes often involve a complex and iterative process and rely on inputs from domain experts as well as data scientists. Algorithms developed for phenotype detection then have to be adapted to the specific local data environments at a particular institution. The implemented algorithms are not easily portable because different institutions may use different data models and different terminologies. An ongoing eMERGE pilot project aims to develop novel strategies for reusable and sharable computational algorithm implementations dealing with different source data models, terminologies and algorithm logics. Our main strategy is to separate the source data environment from algorithm logic. The logical separation has been explored before by the Arden Syntax to share task-specific knowledge implementations across institutions.

Methods
Nineteen eMERGE phenotypes (e.g. atopic dermatitis, ADHD, Asthma, Extreme Obesity, Height, Diabetes/HTN associated CKD, Type 2 Diabetes Mellitus, Dementia, Early Childhood Obesity) that mainly use structured EHR data were implemented as parameterized and modularized queries (implemented query codes are available at https://phekb.org/). The source data are the Columbia University Clinical Data Warehouse in the OMOP Common Data Model (CDM) v4. The parameterization aspect of our query approach deals with different terminologies and source data schemas by using parameter placeholders for codes and schema element names and then injecting corresponding terminology codes or schema/table/field names at execution time. The Modularization aspect is to build complex queries from simple query building blocks, separating concerns in different blocks. The modular query only interacts with the source data model in the first block and constructs a temporary entity–attribute–value (EAV) storage table to store the algorithm required data elements from the source data. Subsequent query building blocks use temporary tables to implement query logic and to store intermediate results.

Results
The temporary EAV table, constructed in the initial query block, includes person id, event date, event concept id, event type, event numerical and string value. Event type refers to a specific data element that is required in the algorithm, for example asthma diagnosis. The key data elements from source databases required for implementing the 19 phenotyping algorithms are analyzed (Table 1). We also specify in which tables in the OMOP4 CDM these key data elements occur. For illustration purposes, an asthma pseudo-algorithm query is depicted in Figure 1.

The logic query blocks implement all required logic in the eMERGE case/control algorithms: e.g., including or excluding specific data elements, calculating counts of eligible data element, checking if patient enrollment status in a series of years is continuous, generating a patient-specific index date from one or more data elements, calculating percentage of qualified data elements, or calculating summary statistics of a data element (e.g. median of height). All logic blocks can retrieve any data element with or without specific time range or relative temporal constraints.

Discussion
Rasmussen et al have explored a modular software approach employing the analytic framework KNIME to share implementations. Different from that, we combine data extraction and algorithmic analysis into a singular query...
based approach. The parameterized and modularized query can isolate core algorithm logic from the source data model. Local terminology coding and database schemata are encapsulated in the parameters. So site-dependent codes can be injected when executing the shared query codes. Consequently, the parameterized and modularized query is sharable and reusable. The pilot project implementations use source data in OMOP4 CDM, but it can also be adapted for other data models by only revising the first block. In addition, separating the query logic from the source data model improves the comprehensibility and reviewability of the query logic. Putting all algorithm related logic into blocks of query code can be challenging and sometimes complicated. For example, finding a patient with continuous enrollment and the median of some measurement. In this pilot project, we only implemented algorithms that require structured data. As a next step, we are exploring the applicability of portable algorithm implementations for phenotypes that use data from unstructured sources.

Table 1. The key data elements from source databases required for implementing the 19 phenotyping algorithms

<table>
<thead>
<tr>
<th>Clinical information</th>
<th>Example Data elements</th>
<th>Table in OMOP4 CDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>person id, birthday, gender, race, ethnicity</td>
<td>person</td>
</tr>
<tr>
<td>Encounter</td>
<td>person id, encounter type (inpatient, outpatient, emergency, etc.), start and end date</td>
<td>visit_occurrence</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>person id, diagnosis code and start date</td>
<td>condition_occurrence</td>
</tr>
<tr>
<td>Medication</td>
<td>person id, medication code and start date</td>
<td>drug_exposure</td>
</tr>
<tr>
<td>Procedure</td>
<td>person id, procedure code and date</td>
<td>procedure_occurrence</td>
</tr>
<tr>
<td>Laboratory and physical exam</td>
<td>person id, lab or physical exam code, value, and date</td>
<td>observation</td>
</tr>
</tbody>
</table>

Parameterized & Modularized Query Example

```sql
USE $databaseName; GO
/* *** Block 1: Extract relevant data element *** */
INSERT INTO #tmp
SELECT $person_id, $condition_start_date AS eventDate, $condition_concept_id AS eventConceptId, 'DxAsthma' AS eventType
FROM $conditionOccurrenceTable CO
WHERE $condition_concept_id IN ($dxAsthmaCd)

/* *** Block 2: Get algorithm related variables *** */
INSERT INTO #AlgVar
SELECT person_id, COUNT(DISTINCT eventDate) AS dxAsthmaCnt
FROM #tmp
WHERE eventType = 'DxAsthma'
GROUP BY person_id

/* *** Block 3: Phenotyping *** */
SELECT WHEN dxAsthmaCnt> 5 THEN CASEphenotype
, WHEN dxAsthmaCnt= 0 THEN CONTROLphenotype
FROM #AlgVar
```

Figure 1. A parameterized and modularized query for pseudo Asthma algorithm and example implementation

Acknowledgments This study was supported by the National Human Genome Research Institute, National Institutes of Health (U01HG008680, PIs: Weng, Gharavi, Hripcsak) and National Library of Medicine (R01LM006910, PI: Hripcsak).

References

Joint Modeling of Survival Events through Multi-task Learning Framework

Zhaonan Sun, PhD\textsuperscript{1}, Xu Liu\textsuperscript{2}, Ping Zhang, PhD\textsuperscript{1}, Jianying Hu, PhD\textsuperscript{1}, Juan Wisnivesky\textsuperscript{3}, MD, DrPH, Fei Wang, PhD\textsuperscript{2}

\textsuperscript{1}IBM T.J. Watson Research Center, Yorktown Heights, NY; \textsuperscript{2}University of Connecticut, Storrs, CT; \textsuperscript{3}Division of General Internal Medicine and Pulmonary Critical Care, Department of Medicine, Mount Sinai School of Medicine, New York, NY

Introduction

Survival analysis, refers to a set of models for analyzing data where the outcome is the time until the occurrence of an event of interest. The event can be death, onset of a disease, or any other events under well-defined circumstances. Over the past few decades, survival analysis has gained wide and successful applications in medical, bioinformatics, epidemiology, and reliability studies. Cox Regression \cite{1} is by far the most extensively used survival model in the application of biomedical research. Classic Cox Regression models one event at a time. In practice, there could exist multiple associated events which are all of interest. Jointly modeling multiple events has the potential to improve modeling performance and gain more insights into the association between the targeting events. The Multi-Task Learning (MTL) approach, which jointly modeling multiple related targets, has recently drawn the attention of many researchers. In this study, we propose a novel approach that incorporates Cox regression into a MTL framework that is able to jointly modeling multiple associated events.

Methods

We used the data from The Cancer Genome Atlas (TCGA) datasets. Specifically, we collected the survival data and the micro RNA (miRNA) expression data for four types of cancer, which are breast cancer, colon cancer, kidney cancer, and lung cancer. Expression level data from 269 miRNAs are available for all four data sets. We used these commonly available miRNAs in this study. The survival of each individual type of cancer is regarded as an event of interest. The sample size (number of patients) for the four data sets are 105, 92, 122, and 106, respectively. Note that the proposed method allows data for different events to be generated from different sets of patients. For each data set, we performed the proportional hazard test, and none of the 269 miRNAs violates the proportional hazard assumption. The rationale for the proposed method is that different types of cancers may share some common mechanisms or genetic factors. The shared mechanisms or genetic factors are regarded as the association between events. If each event is modeled individually, which is referred to as the Single-Task Learning (STL) approach in this study, the association between events are totally ignored. In our proposed method, we intend to utilize the association between events. Let $k$ denote the number of targeting events in a study. In this study, $k$ equals four. In the proposed MTL approach, Cox regression is used as the base model for each individual event. Each base Cox regression model has a coefficient vector, denoted as $\beta_i$ for $i = 1, \ldots, k$, and each element in $\beta_i$ reflect the impact of a genetic factor (i.e. miRNA expression level) on the $i$-th event. If two events share the same genetic factor, the coefficients of the same miRNA in the two base Cox models are likely to be similar. Based on this assumption, the proposed method imposes a hierarchical model on $\beta_i$’s. We used $\Omega$ denotes the hyper-parameter in the imposed hierarchical model. In this study, $\Omega$ is generated from the human disease network \cite{2}. We further impose L-1 regularizer on $\beta_i$’s to enforce sparse solutions. The above mentioned components were finally formed into a convex optimization problem, and estimates of $\beta_i$’s can be obtained by solving the optimization problem. Figure 1 depicts the comparison between STL approach and the proposed MTL approach. The left panel illustrate the STL approach. In this study, we use Cox regression as the STL method, i.e. each event is modeled by Cox regression individually. The right panel in Figure 1 illustrates the structure of the MTL approach.

Results

In this study, we intend to demonstrate the advantage of jointly modeling multiple associated events by comparing the results from STL and MTL approaches. We use the concordance index (C-index) as the comparison criterion. For MTL, we use 5-fold cross validation to select the parameter for L-1 regularizer and to obtain the C-indices. For STL, 5-fold cross validation is used to obtain C-indices for each targeting event. Average C-index from the 5-fold cross validation is used for comparison. The results are summarized in Table 1. The MTL approach outperformed STL approach. The results demonstrated the advantage of taking the event association into consideration.
We further examined the structure of learned $\beta_i$’s. Remind that we impose sparsity assumption on $\beta_i$’s. If the j-th element in $\beta_i$ is non-zero, the j-th feature can be regarded as having certain effect on the i-th event. If the j-th has non-zero coefficient in at least two types of cancers, it can be regarded as a shared factor. On average, 23.5% of $\beta_i$’s are non-zero. We cross checked the shared factors with the miR Cancer database (http://mir cancer.ecu.edu/browse.jsp), many shared factors can be found to be relate to more than one type of cancer. For example, hsa-mir-1 has been reported to be related to breast cancer, kidney cancer, and lung cancer, hsa-mir-143 has been reported to be related to breast cancer, colon cancer, and lung cancer. Due to limited space, we will not report the full list in this article. Further investigation will be reported in our future study.

<table>
<thead>
<tr>
<th></th>
<th>Breast</th>
<th>Colon</th>
<th>Kidney</th>
<th>Lung</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>0.4898</td>
<td>0.4818</td>
<td>0.4992</td>
<td>0.4925</td>
</tr>
<tr>
<td>MTL</td>
<td>0.6946</td>
<td>0.5470</td>
<td>0.5850</td>
<td>0.5612</td>
</tr>
</tbody>
</table>

Table 1. C-index from STL and MTL approaches.
Towards a Formal Account of the Dynamics of Knowledge and Context in Surgical Rooms for the Practice of Surgical Safety CheckLists

Xing Tan, PhD\textsuperscript{1}, Xiangdong An, PhD\textsuperscript{1}, Nicholas Pairaudeau, MD\textsuperscript{2}, Jimmy Huang, PhD\textsuperscript{1}

\textsuperscript{1}York University, Toronto, ON, Canada; \textsuperscript{2}North York General Hospital & Department of Obstetrics and Gynaecology, University of Toronto, Toronto, ON, Canada

**Introduction**

For the purpose of reducing rates of surgical complication and death in hospitals and other medical settings, World Health Organization (WHO) Patient Safety has introduced and has been actively promoting worldwide implementation and usage of its Surgical Safety CheckLists (SSCL)\textsuperscript{1,2}. SSCL compile essential surgical safety objectives and steps into items to be checked by clinicians at work. There are strong evidences to support that, appropriate use of SSCL and adherence to it would lead to significance in reinforcement of accepted surgical safety practices, and in improvement of communication among operating team members\textsuperscript{3,4}. Nevertheless, further improvement and reliable use of SSCL require its formal integration with the pervasive computing systems in today's surgery rooms. In helping clinicians to share contextual knowledge with each other and in enabling context-aware and automated reasoning, the roles of ontologies and action theories are important for integration: An ontology would provide a formal specification of concepts related to SSCL, including people, places, medical instruments, operations events, and etc.; An action theory would offer a computerized mechanism to keep track of the change of operative context, thus provide a formal and on-site scrutinization on the appropriateness of application of steps enlisted in SSCL.

**Methods**

The methodology makes use of two foundational technologies in knowledge representation and reasoning\textsuperscript{5}: ontologies, referring in particular to a controlled vocabulary for SSCL; and action theories, for keeping track of relevant and critical changes of situations and context upon application of actions in surgery rooms. More precisely, we develop an ontology FOSS (Formal Ontology for Surgical Safety) and an action theory ACSS (Action Theory for Surgical Safety) in this research, supporting accurate implementation and pervasive usage of SSCL.

**Ontology**

An ontology\textsuperscript{6,7} comprises of syntactically structured and semantically rigorous descriptions of entities, attributes of entities, and relationships between entities in the domain of interest. When knowledge is represented ontologically, inferences of implication and consequence in the domain can be performed through computer-facilitated reasoning. Description Logics (DL) are logical systems that are widely used in ontological modeling, for manipulating compound predicates such as "A circulating nurse who is assigned as a Checklist coordinator in all operating teams at this hospital". DLs provide a logical foundation for OWL\textsuperscript{8}, a Web Ontology Language as standardised by the World Wide Web Consortium (W3C). OWL ontologies are often developed via Protege\textsuperscript{9}, a GUI-based integrated environment for ontology development.

Design of the ontology FOSS requires that it is competent enough to capture the following categories of surgical room knowledge: (subsumption) objects in the domain such as personnel and equipments are grouped into classes. One class could be subsumed by the other (e.g., the concept Anaesthetist is subsumed by the concept OperationTeamMember). (composition) One class comprises several others (e.g., OperationTeam comprises Surgeon, Anaesthetist, Nurse, Technician, and Other). (role) Tasks are specified as roles. The individuals who perform these tasks are defined as composite concepts (e.g., A ChecklistCoordinator, who in his role performing the task :checkItem, must be either a Nurse or a Surgeon). (membership) An individual/instance is a member of a class. (e.g., nancy is a Nurse).

**Action Theory**

Action theories in formal languages such as Event Calculus\textsuperscript{10} and Situation Calculus\textsuperscript{11,12}, are logic formalisms for representation of actions and reasoning on their effects in dynamical systems. This research uses Situation Calculus. There are two main ingredients in any Situation Calculus-based action theory: 1) Fluent conditions, whose actual values can be updated from applications of relevant actions. 2) Actions, applications of which will bring effects in term of changes of values of certain conditions (i.e., context) in a system. At first, there is an initial situation called S0. Starting from S0, as the system evolves over sequences of actions, the current situation of the system is unfolded into its future situations.
The action theory ATSS should include: A collection of fluent conditions (e.g., marked(X), indicating whether X is marked, where X is a variable and it can be instantiated into a particular site; Or counted(X), indicating whether X is counted and X can be instantiated into particular facilities, instruments, equipments, or items such as sponges and needles). A collection of actions (e.g., mark(X), an action to mark X; and count(X), an action to count X). **Precondition Axioms**, which are formal specifications on the applicability of actions (e.g., before the action inciseSkin, it is required that the site need to marked, marked(site) is true). **Effect Axioms**, which define the effect of actions (e.g., after the application of mark(site1), marked(site1) is set to true).

**Results**

This research presents FOSS and ATSS. Key concepts in FOSS include Place, Agent (human beings, hardware agents or software ones) and Role. **OperatingSuite** is defined as a compound place that comprises of **OperatingRoom, WashRoom, changeRoom, RestRoom, PreparationRoom, Storage, Office, RecoveryRoom, Corridor, and Other. OperatingRoom** hosts **Operation**, which is a sub-class of **Process**. All agents have properties. In the same time, agents can be associated with certain roles.

Our basic thesis is, applicability of actions depends on the context where actions are to be executed, meanwhile, applications of actions change the actual settings of the current context. In ATSS, applicability and effects are captured by Precondition Axioms and Effect Axioms. Each action type in ATSS is assigned one Precondition Axiom and one Effect Axiom. From these basic, atomic actions (ones that cannot be further decomposed), we can define iteratively compositied, more complex actions, till eventually we define the complex action check, which consists sequentially three sub-actions: **signIn, timeOut, signOut**. Any instantiation of the action check indicates one particular realization of the operative procedure that meets accepted safety requirements.

**Discussion**

Context information is critical to the effective use of SSCL in surgical rooms. A formal account (in terms of ontologies and action theories) of knowledge and context information enables integration of SSCL into their computational environments in surgery rooms. In this research, FOSS an OWL ontology to capture the knowledge in surgery rooms for the practice of SSCL is developed. Different from other recent efforts and approaches, we provide axiomatization of dynamics of context, through developing an action theory ATSS, in the formal language of Situation Calculus. We would also consider using upper level ontologies for building up FOSS. We will post our ontologies on [http://www.yorku.ca/adersim/xtan/ontologies](http://www.yorku.ca/adersim/xtan/ontologies)

**Acknowledgement**

This research is supported by the BRAIN Alliance of Ontario Research Fund - Research Excellence (ORF-RE), the CREATE program of Natural Sciences and Engineering Research Council of Canada (NSERC), and Dapasoft. Inc.

**References**

Activated Working Baby Boomers Personal Health Information Management Needs and Expectations: An Exploratory Study

Donghua Tao, PhD, MA, MS1, Deborah E. Seale, PhD1, Cynthia M. LeRouge, PhD2
Jennifer Ohs, PhD1, Helen W. Lach, PhD1, Keri Jupka, MPH1, Ricardo Wray, PhD, MS1
1 Saint Louis University, Saint Louis, MO
2 University of Washington, Seattle, WA

Introduction

Baby boomers (adults born between the years of 1946 and 1964) make up the largest segment of the population in many countries, including the United States (about 26%)1. While they have increasing medical needs and thus demand more health care resources when they reach retirement age and beyond, they appear to understand the need to be proactive in order to achieve long-term health and successful aging. As a “sandwich generation”, baby boomers tend to have competing demands of work and family care, and they tend to be caring for both children and aging parents. Baby boomers, with more education and higher level jobs, and being witnesses and early users of communication and information technologies, are likely to want to take responsibility for managing their health and participate in health-related decision making (i.e., activated working baby boomers). To effectively take on these responsibilities, baby boomers need to efficiently manage and use their personal health information (PHI) (e.g., lab results, doctor contacts, medication lists, health insurance information, and information searched online, etc.)2. As a result, the need for better tools to support PHI management will increase substantially. This study investigated activated working baby boomers’ PHI management needs and expectations through focus groups.

Methods

Focus group interviews were used for the data collection with the IRB approval. Working baby boomers who were born between 1946-1964, full-time or part-time employed, and self-identified in a screening survey as being active in managing their health were the study participants. A two-step purposive sampling technique was used. First, a metropolitan non-profit business coalition was contacted to identify the organizations who agreed to facilitate in recruiting their employees for the study. Second, an email with a link to a brief screening and demographic survey was sent to the employees of those identified organizations. Eligible employees were those who answered positive in their response to 4 out of 5 following screening questions: 1) want to be actively involved in decision-making about my health; 2) see a healthcare provider for a routine checkup annually; 3) regularly look for health information to understand or manage my health; 4) use a computer, tablet or smart phone on a weekly basis; and 5) are interested in using technology to manage my health. Of 210 employees who completed the survey, 155 (73%) met eligibility criteria and received the focus group invitations, and 57 attended with age between 51 to 69 years old (mean = 57), primarily married white female, had at least a college degree with the income of at least $75,000 a year.

Six focus groups were conducted after work, at the work site, with dinner and a small monetary incentive provided. Each session was about 90 minutes long and audio-recorded. An experienced moderator who was not part of the research team led the discussions using a standard interview guide. Participants chose pseudonyms to ensure that no personal identifiers were recorded. The Gagnon and Chartier Citizen-Patient 3.0 Profile3 was used as a conceptual framework to develop focus group questions and guide transcripts coding: 1) search appropriate and secured medical information about his case and improve his “health literacy”; and 2) use consumer self-management health technology in a secured manner to manage his health. Focus group transcripts were coded with Dedoose, qualitative software. Transcripts were coded independently by three teams with two members in each team using theoretical thematic analysis4. Each team met to discuss discrepancies and came to agreement on coding.

Results

Three significant themes emerged related to activated working baby boomers PHI management behaviors, needs and expectations to better understand and manage their health. Working baby boomers are: 1) active searchers and users of health information on the Internet; 2) aware of the variability in online health information quality; 3) are challenged to access, organize and share PHI of their own and their family members.

Working baby boomers are Internet savvy
As activated healthcare consumers, working baby boomers in this study have the willingness and ability to search and use appropriate and trusted health information. Acting as their own “agent”, they want to be able to search, choose, and use health information to: 1) understand conditions, treatment and care; 2) compare perspectives and options concerning conditions, treatments and care in order to make health-related decisions, such as whether or not to have surgery, choose doctors, etc.; and 3) prevent onset or worsening of a condition, as well as maintain their overall health. Internet (e.g. Google) seems to be a default place for them to search for information about health, although health care providers, relatives and friends, and even unknown individuals from social media sites are also the sources for obtaining health information.

**Awareness of the variability in online health information quality**

Working baby boomers in this study are cognizant of information overload and the importance of the accuracy of medical information. They intend to and also have the ability to sift through online health information and identify trusted information to manage their health. They search health information from known reliable information resources, such as Mayo Clinic. They also make effort to verify specific health information by comparing multiple resources, including professional websites, for the same health-related topics or issues. However, they found it difficult to handle conflicting information and felt “a challenge and a risk in going to the Internet”. They wish there was “a way for the Internet to be a more reliable source of medical information” that “could be like librarians who have access to... who could tell you about landscape or tell you about any of other things that you may be unaware of...” There seems to be a gap between baby boomers’ willingness and the ability to appraise and apply health information to their situations and the availability and/or awareness of reliable information resources.

**Challenge of organizing and sharing personal health information**

Working baby boomers in this study found it challenging to organize and share their PHI. They repeatedly expressed the need to “set up my own health file” and keep their own health notes, test results and medication lists, etc., in order to track their health history and monitor changes in their health and medical conditions over time. However, their PHI was fragmented and scattered in multiple locations, repositories, devices or applications in both paper and electronic format. Many of them said they had difficulty in capturing, remembering and understanding what the doctor said and did during a visit. They had trouble knowing “how it’s spelled” and “you feel silly asking.” They expect doctors can “teach the patients how to set up their own personal health files” so “they’ve got a track record.”

This issue is especially a challenge for working baby boomers as a “sandwich generation”. Patient portal or Personal Health Record (PHR) systems seem to serve as a central place to store PHI. However, working baby boomers do not have control on what information they can access and input in most patient portals or PHR systems available today. Moreover, none of systems hold all PHI for the same patient across different health care systems.

Fragmented PHI creates difficulties in sharing PHI between providers and working baby boomers. They want to share their information with doctors of their choice, and “with the emergency room doctor that’s not in the network” and “link (their PHI) to specialists outside that group”. In order to do that, baby boomers kept separate lists in their wallets, retyped or printed out or copied down information from their portal to fill out physician’s forms. Participants wanted a consumer health technology that would allow them to share their PHI with whomever they want. They suggested having a chip or a thumb drive for updating and transferring PHI with health care team; or an app for the smart phone that can “sync the data that I have on my smart phone to their system” and vice versa.

**Discussion**

Activated working baby boomers care about their health and also want to be heavily involved in their health-related decision making and problem solving. Health information and knowledge empower them to take responsibility for managing their health and user-centered designed technology can help them successfully take on these responsibilities. A tool that can provide evaluated online health information and an app or a system that allows activated working baby boomers to obtain and share relevant PHI with whoever they want are two possible technology solutions for meeting the critical needs of working baby boomers in accessing reliable health information and effectively organizing and sharing their PHI.

**References**

Evaluating content coverage of a dental diagnostic terminology

H. Taylor, MPH, LDH¹, Z. Siddiqui, BDS, MS¹, K. Frazier, BS¹, T. Thyvalikakath, DMD, MDS, PhD¹,²
1- Indiana University School of Dentistry, Indianapolis, IN
2- Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN

Introduction:

Dental diagnostic terminologies can be helpful in facilitating health information exchange, evaluating clinical outcomes, dental disease prognosis, evidence of appropriateness of care, identifying best practices, and improving diagnostic skills of students and clinical research¹. However, unlike medicine, there is no commonly accepted diagnostic terminology in dentistry². Currently, two dental diagnostic terminologies are available for use by clinicians and educators – Systematized Nomenclature of Dentistry (SNODENT) and Dental Diagnostic System (DDS). Both SNODENT and DDS Codes are not widely used in dental practices³; however a few dental schools have implemented the use of either SNODENT or DDS Codes into their clinical training sessions with dental students. Little data exists on the strengths, weaknesses, and usefulness of SNODENT, particularly from the perspective of education of faculty and students. This particular study reports on evaluating the content coverage of SNODENT within the Electronic Dental Record (EDR) used at the Indiana University School of Dentistry (IUSD).

Methods:

This study was approved by the Indiana University Institutional Review Board (Study #: 1503902699) and conducted at the IUSD. A convenience sample of 20 participants consisting of faculty, third-year dental students (DS 3), and fourth-year dental students (DS 4) of IUSD participated in the study. Participants were requested to evaluate a hypothetical dental case and asked to “think aloud” and document on paper their diagnosis for specific dental conditions as they reviewed the study patient’s medical history, dental findings, photographs, and radiographs. This hypothetical patient had clinical findings commonly encountered at the IUSD clinics. Participants were then asked to work within the treatment planning module of the Electronic Dental Record (EDR) (axiUm, Exan corporation, Vancouver, BC, Canada), and select the best SNODENT diagnostic code(s) that represented their written diagnosis. Subsequently, participants determined if the selected code(s) matched their written diagnosis “completely”, “partially”, or “not at all”. At the end of the study session, each participant completed a questionnaire based on their experience using SNODENT. The questionnaire had four questions on a four point Likert scale that ranged from strongly agree to strongly disagree (See Table 1).

Participants’ interaction with the EDR was recorded using screen and voice capture software (Camtasia®). Two investigators independently compared the participants' written diagnosis and the diagnosis code they selected in SNODENT to determine the match between the two diagnosis. Participants were classified into specific groups (DS 3, DS 4, and faculty). Sensitivity testing was performed to detect changes between participants’ written diagnosis and SNODENT code(s) selection. Chi-square test was performed to detect any significant difference among the groups. Participants' satisfaction with SNODENT was summarized using descriptive statistics. Cochran's-Mantel-Haenszel (CMH) test was performed to assess any consistent difference in participants' satisfaction among the three groups. The software IBM SPSS Statistics Version 23 (SPSS, Inc., Chicago, IL, USA) was used to perform the statistical analysis. Both Chi-square test and CMH test was performed at 95% confidence interval and 5% significance level. We analyzed participants' response to questionnaire using descriptive statistics.

Results:

The 20 participants selected a total of 251 SNODENT codes to diagnose the dental conditions of the hypothetical dental patient. The DS 3 and DS 4 students demonstrated high match between their written diagnosis and the SNODENT diagnosis they selected with sensitivity levels of 74.6% and 74.3% respectively. We observed a low match (sensitivity: 43.9%) between faculty participants’ written diagnosis and the SNODENT codes they selected. Most faculty participants wrote down treatment for the dental conditions instead of the diagnosis. The average sensitivity for the three groups was 64.5%. The three groups were significantly different in the match between their written diagnosis and the selected SNODENT codes, especially between the DS 3 students and the faculty
(p<0.0001) and the DS 4 students and the faculty (p<0.003). However no statistically significant difference was observed between DS 3 and DS 4 students (p 0.97). Participants noted that the following searches found no codes to match their diagnosis: Interproximal primary caries; generalized mild chronic periodontitis; Pulp Polyp; Gross caries; Poor oral hygiene; Non-restorable tooth; and Plaque-induced gingival diseases. The participants were satisfied with 62% of their selection of the SNODENT terminologies, partially satisfied with 33% and dissatisfied with 5% of the SNODENT terminologies. No significant differences (p 0.258) were observed among the three groups’ satisfaction with their SNODENT selections. As shown in Table 1, 19 out of the 20 participants strongly or mildly agreed that the SNODENT terminology worked well with entering appropriate diagnosis and will add value to patient care and treatment planning. Fifteen participants agreed that it is easy to understand and navigate through the SNODENT system and 12 participants agreed it is easy to implement use of SNODENT in routine patient care.

Table 1: Participants’ response to questionnaire related to SNODENT diagnostic terminology

<table>
<thead>
<tr>
<th>Questions</th>
<th>Strongly Agree</th>
<th>Mildly Agree</th>
<th>Slightly Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: This coding tool worked well in entering appropriate diagnoses for this patient's unique case.</td>
<td>6 (30%)</td>
<td>13 (65%)</td>
<td>0 (0%)</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>Q2: I believe using this coding tool will add value to patient care and treatment planning.</td>
<td>9 (45%)</td>
<td>10 (50%)</td>
<td>1 (5%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Q3: I found this coding tool easy to navigate and understand.</td>
<td>6 (30%)</td>
<td>9 (45%)</td>
<td>3 (15%)</td>
<td>2 (10%)</td>
</tr>
<tr>
<td>Q4: I believe this coding tool would be easy to implement in daily clinical care activities.</td>
<td>6 (30%)</td>
<td>6 (30%)</td>
<td>6 (30%)</td>
<td>2 (10%)</td>
</tr>
</tbody>
</table>

Discussion:

Dental students selected SNODENT codes that represented their written diagnosis at higher sensitivity levels than the faculty (DS 3: 74.6%; DS 4: 74.3%; Faculty: 43.9%). Interestingly, when study investigators analyzed faculty’s original written diagnosis for certain teeth, faculty wrote down treatment for these teeth as opposed to diagnosis. Faculty’s tendency to consider a patient case in the constructs of dental treatment rather than diagnosis is noteworthy and common in seasoned dental clinicians. This fact lends explanation to the low sensitivity levels of faculty’s selection of SNODENT codes compared to their written diagnosis.

The majority of SNODENT codes selected by all participants were sufficient in representing their original written diagnosis for selected teeth. However, a number of participants felt “frustrated” at their inability to locate a code that best represented their original diagnosis. The participants’ overall agreement that SNODENT enables appropriate diagnosis entry and adds value to patient care indicates the potential of diagnostic terminologies to improve patient care. However mixed responses to the ease of use and implementation of this terminology in daily practice shows there is a learning curve with the system and like any system in clinical practices, it will take some time to be commonly accepted by clinicians.

Conclusion:

Education on dental diagnosis is needed for dental students and faculty, especially faculty who are trained to think in terms of treatment for each patient as they work through a case. Coding term additions to the SNODENT coding terminology could improve satisfaction levels of dental practitioners’ use of the SNODENT system.

References:

Convening Utah’s Healthcare Interests – A Statewide Care Coordination Infrastructure

Sidney N. Thornton, PhD1,2, Shan He, PhD1, Iona Thraen, PhD2,3, Matt Hoffman, MD4, Darren K. Mann1, Wu Xu, PhD2,3

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

Introduction

Care of patients is dispersed across diverse health care professionals from different organizations1,2. Lack of timely information transfer among these providers can lead to medication errors, duplicative tests, and lack of follow-up care3. To enhance coordinated care, a community care team and supporting infrastructure need to be built beyond enterprise-wide care integration efforts and point-to-point connections. As the adoption of Electronic Health Record (EHR) systems in U.S. increases and interoperability among disparate systems enhances, the time is opportune to promote an infrastructure for care coordination among health professionals along the care continuum. The objectives of this project were to a) build a statewide infrastructure for care coordination across healthcare, public health systems and consumers, b) demonstrate automated care coordination workflows, and c) report our experiences and lessons learned including technical and non-technical strategies, opportunities and barriers.

Methods

In pursuit of a statewide solution for care coordination, stakeholders in the Utah healthcare industry including the Utah Health Information Network (UHIN) that governs the state-designated clinical Health Information Exchange—the cHIE, public and private healthcare providers (e.g., University of Utah Healthcare, Intermountain Healthcare, Revere Health, etc.), HealthInsight—a quality improvement organization, and Utah Department of Health (UDOH) convened in the form of governance, service, technical and legal committees to work together on solving emerging needs for coordinated care across the community. Instead of building expensive and independent organizational solutions, a statewide care coordination infrastructure leveraging existing resources was conceived as the most pragmatic approach.

First, as prioritized by the resource investment by the participating stakeholders, five cross-system care coordination use cases were selected for analysis to determine the core components of the care coordination infrastructure: 1) Newborn care bundle; 2) Discharge to skilled nursing facility (SNF); 3) Collaboration between the Utah Poison Control Center and emergency departments on poison exposure cases; 4) Physician Orders for Life Sustaining Treatment form and care summary access by Emergency Medical Services; and 5) Physical and behavioral health integration. Each use case represents a real-world scenario where multiple healthcare professionals across different systems are involved in a patient’s care. The data flow analysis showed that major care coordination components (e.g., message sending and receiving, event listeners, patient record linkage, community-based care team identification, workflow management, patient alert/notification, etc.) are common functionalities across use cases. The information transfer flow and connection requirement for each use case was analyzed in detail. One such example is the newborn use case, where the care continuum spans across prenatal care by an obstetrician, newborn care and first blood spot specimen collection at the birthing facility, newborn screening test at the state laboratory, newborn hearing screening by an audiologist, well-child visit and second blood spot specimen collection at a pediatrics clinic, public health follow-up programs for newborns with abnormal screening results, genetic consultation and specialty care. Regulatory record submissions related to newborns such as vital records are also required by UDOH. The length of the time span and the number of stakeholders involved in a newborn’s care make it impossible for any single system to support a smooth, connected care experience.

Second, existing local resources and interoperability capacities were evaluated to assess which modules can be integrated into the new care coordination infrastructure. Existing data resources include organizational and statewide master patient indices, organizational provider directories, clinical data repositories at the cHIE, billing records, all payers claim database, and public health vital records. Existing system and connection resources include certified EHR adoption and patient portals provided by major healthcare providers, and standard connections in operation among major participants.
Finally, results from the use case analysis and resource evaluation were consolidated into the care coordination infrastructure. Three use cases were implemented as pilot projects to evaluate and refine the infrastructure design.

**Results**

The Utah care coordination infrastructure leverages existing central repositories and HIE capabilities in the community in order to achieve automated information transfer across healthcare, public health systems as well as patients and caregivers. The key components of the care coordination infrastructure are illustrated in Figure 1.

![Figure 1 Utah Statewide Care Coordination Infrastructure](image)

The Care Coordination Broker (CCB) is the central logic engine that determines what information needs to be sent to whom upon what event and through which mechanism. The business process modeling standard—Business Process Management Notation (BPMN) is used as the underlying framework in the pilot implementation. It provides a scalable infrastructure to support new use cases in the future. The identity services and patient-provider relationship manager that maintains the community care team for an individual patient are the foundational services invoked by the CCB. Patient notification may be a critical piece for some coordinated care workflows through patient portals and patient-registered devices. The audit service will provide data for future evaluation on the impact of the automated care coordination workflows. The newborn care bundle, poison case management, and advanced discharge planning for SNF transfer are pilot implementations that demonstrate the feasibility of using pre-existing community resources. As more care coordination scenarios emerge and transaction volumes increase, the statewide infrastructure shows promise of efficiency and economy.

**Conclusion**

The collaborative demonstrated an innovative approach to facilitating care coordination by implementing automated information transfer rules on a business process management platform. Existing health information infrastructure in Utah is re-organized and leveraged to support the statewide care coordination services. It is critical to convene major stakeholders and promote collaborative partnership across the community. Creating a statewide care coordination infrastructure to manage patients across the care continuum is essential to achieving optimal health outcome, reducing costs, and enhancing the patient care experience.

**References**

Expert Recommendations on Redesigning Drug Allergy Alerts in Electronic Health Record Systems.

Maxim Topaz RN, PhD1,2*, Foster Goss DO3*, Kimberly Blumenthal MD4,5, Kenneth Lai MSc1, Diane L Seger RPh1,5, Sarah P. Slight PhD, PGDip1,6, Paige G Wickner MD1, George A. Robinson, RPh7, Kin Wah Fung, MD, MS, MA8, Robert C. McClure, MD9, Shelly Spiro, RPh, FASCP10, Warren W Acker, BA1,5, David W. Bates MD, MSc1,2, Li Zhou MD, PhD1,2,3

1General Internal Medicine & Primary Care, Brigham and Women's Hospital, Boston, MA, USA; 2Harvard Medical School, Boston, MA, USA; 3University of Colorado, Department of Emergency Medicine, Aurora, CO, USA; 4Division of Rheumatology, Allergy and Immunology, and Medical Practice Evaluation Center, Department of Medicine, Massachusetts General Hospital, Boston MA; 5Clinical & Quality Analysis, Partners Healthcare System, Wellesley, MA, USA; 6Division of Pharmacy, School of Medicines, Pharmacy and Health, Durham University, Durham, U.K.; 7First Databank, Inc., San Francisco, CA, USA; 8National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; 9MD Partners, Inc., Lafayette, CO, USA; 10Pharmacy HIT Collaborative, USA.*Co-first authorship.

Abstract

Providers are increasingly overriding drug-allergy alerts in electronic health records with average override rate of about one in ten alerts. We present a set of conceptual and practical recommendations to develop a new generation of drug-allergy alert systems. Recommendations were generated by a multidisciplinary group of experts from medicine, pharmacology, nursing, quality and safety, and clinical informatics.

Introduction

Drug allergy alerts are one of the key features of computerized provider order entry (CPOE) in electronic health record systems. These alerts can protect patients and providers from administration of medications that could result in an adverse reaction by comparing ingredients and cross reactivity between prescribed medications and medications on the patient’s allergy list. In practice, these alerts as currently used in most systems have serious limitations with over 90% of these alerts are being overridden.1–6 Observations from the last 15 years show override rates increasing from 50% in the mid 1990s1 to almost 90% in 2015.2,3 Many current alerts are inconsequential and almost never result in an adverse reaction.4,5 It is estimated that providers need to review more than 123 unnecessary alerts to prevent one adverse drug event.4 This presents a substantial problem with drug allergy alert systems that demands redesign. While most of the overrides may simply indicate low value alerts that are not providing useful decision support to providers, many alerts are overridden in situations that are not safe, for example, when patients have a history of severe or immune-mediated reactions. Though surprising, alert override rates for anaphylaxis or angioedema are often greater than 75%.2–6

Methods

New approaches to allergy alerting are critically needed. During 2015, we assembled a multidisciplinary group of experts from internal medicine, allergy/immunology, emergency medicine, pharmacology, pharmacy, nursing, quality and safety, and clinical informatics to identify new methods for drug allergy alerting systems. The group produced a set of conceptual and practical recommendations to develop a new generation of drug-allergy alert systems.

Results

Patient engagement in the allergy reconciliation process is a key to creation and maintenance of meaningful allergy lists. In many cases, allergy is never removed from the patient’s list once it is recorded. At the same time, even when patients have a medication allergy, it may not appear on the list. Strategies to engage patients in reviewing their
allergy information and reconciling allergies with their provider should be further tested and implemented. For example, patients should be able to update their allergy information in personal health portals, which should prompt the clinician to review the information with them on their next encounter.

Improved allergy documentation and characterization of allergic information to improve alert accuracy. It is crucial for allergy information to accurately characterize and distinguish adverse drug reactions as side effects, toxicities, intolerance, idiosyncrasy, or allergies. Clinicians recording these events may not know the distinctions and computerized clinical decision support could assist in these determinations. A more detailed specification of the patient’s allergy at the time of entry or reconciliation will ensure that alerts are triggered when they matter most, and avoid unnecessary alerts on mild intolerances or previously-tolerated medications.

Alerting mechanism: Allergy alerting systems should consider reaction severity and other contextual information (e.g., the type of match between the allergen and prescribed medication, reaction occurrence probabilities, information on whether this alert was fired or overridden in the past, etc.) into consideration when presenting alerts to clinicians. Fuzzy and disorganized alerting approaches produce noisy alerts that are not clinically meaningful. One potential solution that is common in drug-drug interaction alerts is alert tiering. Based on adverse drug reaction severity, or the likelihood of an immune-mediated reactions, such as immediate, IgE-mediated allergic reaction that on repeat exposure could lead to severe outcomes, drug allergy alerts can be classified as either informative (providers will still see the alert information but would not need to take an action to override the alert) or interruptive (providers will be required to provide a reason for override or cancel the prescription).

Continuous alert monitoring and improvement: organizations should track their allergy alerting and override rates over time. This will help identify changes in alerting patterns and turn-off alerts that are disruptive.

Discussion

We believe that these recommendations can advance and improve the current state of allergy alerting. Patient safety will improve if providers are not being barraged with inconsequential alerts, making them more likely to adhere to warnings that represent a serious concern.

References


Novel Approaches to Medication Teaching for Complex Medication Regimens

Demetra Tsapepas, PharmD1,2, Hojjat Salsamian, MD MPH PhD1,2, Sumit Mohan MD, MPH1,2, Jennifer Prey, MPhil MS2, Andrea Wells-Shapiro CPXP1, David K Vawdrey, PhD1,2

1NewYork-Presbyterian Hospital, New York, NY; 2Columbia University, New York, NY

Introduction

Mobile applications can enhance healthcare providers’ ability to deliver tailored interactive patient education at the bedside, but the use of such applications is not well understood, particularly in specialized populations.1,3 Solid organ transplant recipients, for instance, have complex medication therapy regimens requiring extensive personalized education. We evaluated the impact of a digital education program delivered through a mobile application on transplant patients’ awareness of their medication regimen and associated side effects.

Methods

This single-center retrospective study of adult kidney or pancreas allograft recipients transplanted from January 2015 to June 2016 analyzed the results of a survey distributed among patients who were educated through a novel teaching approach using tablet computers. Using interactive content on the tablet computer, hospitalized patients engaged in their own learning about their immunosuppressive regimen; reviewed the name, indication, dose, side effects, interactions, and special considerations for each of medication. Concise, stimulating video clips about each medicine were designed to appeal to a broad patient audience including those with limited healthcare literacy. Video clips were two to eight minutes in duration and consisted of images, text, and voice-over (see Figure). To reinforce key concepts and identify knowledge gaps, the application contained assessment questions followed by feedback that included explanations of the correct answers. Each patient answered 19 assessment questions, 14 focused on the medication indication and important information and five questions focused on side effects. Patients were encouraged to spend as much time as they needed to complete the module. A tailored in-person follow-up session was also conducted by a pharmacist to reinforce content and address knowledge gaps.

Results

Survey respondents included 183 patients who used the virtual teaching tool. Patients were ethnically diverse, ranged in age from 21–75 years, and 25% had an undergraduate degree or some college education. Ninety percent of patients were educated on our standard dual-maintenance immunosuppressive regimen and three opportunistic infection prophylaxis agents. While the majority of patients (90%) were able to identify the correct answers for questions related to drug indication, dosing, and special rules, many (62%) had difficulty recalling information about side effects (see Figure). The majority of patients (90%) favorably ranked this digital education program.

Conclusion

Utilizing virtual teaching via mobile applications may improve the efficient healthcare delivery by providing consistent, quality patient education interventions. In order to succeed, patient centered care initiatives need to empower patients to actively participate in their care. To our knowledge, this is one of the first and largest studies to demonstrate the efficacy of structured education program for transplant patients with complex medication regimens.

References

Diagnostic Journeys of Patients Evaluated for Lyme Disease and Given Extended Antibiotic Therapy

Yi-Ju Tseng, PhD1,2, Alfred DeMaria Jr MD3, Donald A. Goldmann MD4,5, Kenneth D. Mandl, MD, MPH1,6,7

1 Computational Health Informatics Program, Boston Children’s Hospital, MA, USA; 2 Department of Information Management, Chang-Gung University, Taiwan; 3 Bureau of Infectious Disease and Laboratory Sciences, Massachusetts Department of Public Health, MA, USA; 4 Institute for Healthcare Improvement, MA, USA; 5 Division of Infectious Diseases, Boston Children's Hospital, MA, USA; 6 Department of Pediatrics, Harvard Medical School, MA, USA; 7 Department of Biomedical Informatics, Harvard Medical School, MA, USA

Introduction
Lyne disease (LD) is most straightforwardly diagnosed when erythema migrans (EM) is observed, there is a history of possible exposure to ticks, or a serologic test is confirmatory. However, the diagnosis can be far from straightforward, particularly if the initial diagnosis isn’t made because EM doesn’t develop or isn’t noticed, the patient is unaware of tick exposure or bite, or the presentation is atypical. LD can be treated effectively with 2-4 weeks of antibiotics. If untreated, LD may lead to neurologic and rheumatic manifestations weeks to months later. Even after standard treatment, about 10-20% of patients with LD report nonspecific persistent symptoms, may result in a diagnosis of post-treatment LD syndrome, especially when the diagnosis is delayed or the treatment is inadequate. We sought to map diagnostic journeys taken by patients diagnosed with LD, hypothesizing that those ultimately receiving extended antibiotics may be clinically distinct.

Methods
We performed a population-based retrospective cohort study using insurance claims from a nationwide, private health insurance plan in the US from 2010-2012, for patients residing in the 14 states with a high-prevalence of LD—Connecticut, Delaware, Maine, Maryland, Massachusetts, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Virginia and Wisconsin. The Boston Children’s Hospital Institutional Review Board approved the study, granting a waiver of consent.

We use a published LD claims data-based case definition including LD diagnosis code, serologic test, and antibiotic therapy for LD. Patients evaluated for LD and given standard antibiotic therapy (PLDSA) are defined as those receiving 2 to 5 weeks of antibiotics. Patients evaluated for LD and given extended antibiotic therapy (PLDEA) are defined as receiving ≥5 weeks. To enrich the cohort for patients beginning their diagnostic journeys, we included beneficiaries enrolled in the plan for at least 180 days before an LD diagnosis and who received the first LD diagnosis in 2011 or 2012.

We assessed the association between PLDEA and diagnoses grouped into clinically meaningful diagnostic categories using the Agency for Healthcare Research and Quality Clinical Classifications Software (CCS). Each category is defined as a binary variable, based on the presence or absence of at least one ICD-9 code from the appropriate category occurring during defined 30 day time periods between 1 and 180 before a patient’s first LD diagnosis. The presence of a diagnosis within the diagnostic categories among PLDEA was compared with the risk in PLDSA using logistic regression, adjusted by age and gender, and expressed as odds ratios (ORs). All analyses were performed using R software. All statistical tests were two-sided.

Results
A total of 7,928,693 insured individuals resided in the 14 states with a high-prevalence of LD. Of these, 3,207 (40.45 per 100,000 insured) and 600 (7.57) individuals met the case definitions for PLDSA and PLDEA, respectively. Compared with PLDSA, PLDEA cases were more likely to be female (62.1% vs. 48.5%, P<0.001) and older (42.5 vs. 40.7, P<0.05). Figure 1 maps out the temporal, claims associated diagnostic journeys of PLDEA compared with PLDSA in the six months leading up to the first diagnosis of LD. From 151-180 days before the first LD diagnosis, the risk for claims associated with back problems (OR, 2.1; 95% CI, 1.4-2.9; P < .001) and other connective tissue
disease (1.6; 1.1-2.3; P < .01) were elevated among PLDEA. From 61-90 days before, claims associated with malaise and fatigue (1.7; 1.1-2.6; P < .05), other nervous system disorders (2.0; 1.3-3.1; P < .01), and non-traumatic joint disorders (1.4; 1.0-2.0; P < .05) are more likely among PLDEA. From 1-30 days before, the risks of mental illness (1.6; 1.1-2.0; P < .01) and headache (1.5; 1.1-2.0; P < .05) associated claims are higher. In contrast, the only low risk claims associated diagnostic category in PLDEA is skin and subcutaneous disorders (0.6; 0.5-0.9; P < .01) occurring up to 30 days before the first LD diagnosis in the study period.

![Figure 1](https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp) Claims associated diagnostic journeys. The colors in blocks indicate the odds of diagnostic categories in PLDEA compared with PLDSA, for statistically significant (P<0.05) relationships. The numbers on the color blocks indicate the percentage of PLDEA patients assigned the diagnostic category.

**Discussion**

PLDEA have distinct signatures with evidence of common progressions through alternate diagnoses and recognition of various co-morbidities. They had a higher risk than PLDSA for claims associated nonspecific signs and symptoms consistent with LD including fatigue, headache, joint pain and connective tissue diseases before their first LD diagnosis during the study period. Skin and subcutaneous tissue diagnoses are in the only category more likely to be found among PLDEA. Given the distinct character of the EM rash and the effectiveness of prompt treatment, this is not surprising. Early identification of these individuals based on patterns of diagnosis could reduce the time to diagnosis and yield cost savings associated with misdiagnosis or delayed diagnosis.

**References**

The Role of Culture and Personality in Health Information Communication with Social Network Members

Rupa S. Valdez, PhD1, Thomas M. Guterbock, PhD1, Kara Fitzgibbon, MS1, Ishan C. Williams, PhD1, Hannah K. Menefee, MPH2, Claire A. Wellbeloved-Stone, MPH1
1University of Virginia, Charlottesville, VA; 2Virginia Tech, Blacksburg, VA

Introduction
Consumer health IT solutions are proliferating as support mechanisms for patients who must increasingly take on self-management responsibilities. One aspect of self-management that has received limited attention from consumer health IT designers, but has been documented as burdensome for patients1, is the communication of health information to social network members (i.e., family, friends, acquaintances). These individuals are often engaged in supporting patients with the emotional, instrumental, and spiritual aspects of living with a chronic condition2,3. We sought to understand patient needs for consumer health IT by assessing patients’ existing health information communication (HIC) practices on social media. Specifically, we sought to determine the ways in which individuals with type 2 diabetes communicate health information with members of their social network on Facebook, the largest social media platform4. In this presentation, we report on how aspects of personality and culture were associated with different approaches to HIC with Facebook social networks (communication with Facebook networks both on and off Facebook). Such a study informs design and implementation of culturally informed consumer health IT solutions.

Methods
Our study consisted of three phases: 1) qualitative interviews and survey development (based on qualitative findings), 2) survey piloting and revision, and 3) large-scale sample survey. Eligible participants for all phases were 18 or older, used Facebook, had a diagnosis of type 2 diabetes, and were citizens or residents of the United States. **Phase one:** Twenty-five individuals recruited on Facebook5 participated in phase one. Phase one participants were interviewed about their HIC practices on and off Facebook as well as factors they believed shaped these practices. Relevant to the analysis reported in this abstract, participants were asked about cultural values influencing their approach to HIC. When queried about other factors beyond those explicitly raised by interviewers, participants mentioned the role of personality. To translate these findings into a survey, we conducted a qualitative content analysis of the cultural values mentioned by participants and designed a survey item reflective of these findings. To account for personality, we added the Big Five Personality Test6 to the survey. **Phase two:** Thirteen individuals recruited on Facebook participated in phase two. Phase two participants completed a draft survey based upon phase one results and participated in focus groups, during which they were asked to explain their interpretation of each survey question and discuss ways of altering the wording and organization to limit misinterpretation. **Phase three:** Seven hundred individuals recruited through a commercial survey access panel participated in phase three. Phase three participants completed a finalized version of the online survey, which contained between 75 and 80 questions (depending on version received) and took an average of 22.5 minutes to complete. Relevant to the analysis reported in this abstract, an exploratory factor analysis (EFA) on normalized survey findings (N=650) related to cultural values (Principal Component Analysis as an extraction method and Oblimin with Kaiser Normalization as a rotation method) was conducted to reduce the 29 themes generated by the qualitative content analysis into four factors. Based on an earlier analysis of phase three survey questions about HIC practices, we identified seven approaches to communicating health information with Facebook social networks (Table 1)7. We compared the cultural factors and personality traits reported by survey participants in each profile using one way ANOVA with post-hoc pairwise comparison and Tukey’s HSD method to test for significant differences between profiles (alpha=0.05). The University of Virginia’s Social and Behavioral Sciences IRB approved this study.

Table 1. Seven Profiles: Approaches to general Facebook use and HIC on and off Facebook

<table>
<thead>
<tr>
<th>Category</th>
<th>High General Facebook Use</th>
<th>Low General Facebook Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Health Communicators:</td>
<td>FC-H, n = 161</td>
<td></td>
</tr>
<tr>
<td>High HIC on and off Facebook</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some HIC on and off Facebook</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Off Facebook Only Health Communicators:</td>
<td>OC-H, n = 44</td>
<td>OC-L, n = 84</td>
</tr>
<tr>
<td>No HIC on, and some HIC off Facebook</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Disclosing Health Communicators:</td>
<td>NC-H, n = 26</td>
<td>NC-L, n = 49</td>
</tr>
<tr>
<td>No HIC on or off Facebook</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results
The gender of phase one participants was approximately evenly divided and the gender of phase three participants was unequally distributed with approximately two-thirds of participants identifying as female. The majority of participants in both phases were 30-64 years of age, with about 20% above of the age of 65. In both phases, racial minorities
comprised approximately 40% of the sample. Ethnic minorities comprised 12% of phase one participants and 21% of phase three participants. The results of the exploratory factor analysis of cultural values indicate four underlying factors, which we conceptualized as 1) virtue, 2) comfort, 3) culture, and 4) religion and spirituality (see Table 2). Statistically significant differences among a subset of profiles (α = .05) were demonstrated across all four cultural factors and three of five personality characteristics (see Table 3). Statistically significant differences are indicated by a superscript. For example, there is a statistically significant difference between FC-H and MC-L on the personality characteristic “agreeableness.”

Table 2. Four cultural factors derived from EFA of 29 inductively derived cultural values

<table>
<thead>
<tr>
<th>Factor</th>
<th>Values (Loading)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtue</td>
<td>Work ethic (.699), Open-minded (.788), Education (.849), Equality (.639), Ethics (.856), Family (.785), Individual Liberties (.790), Civil Rights (.806), Positivity (.714), Privacy (.801), Realism (.533), Retirement (.556)</td>
</tr>
<tr>
<td>Comfort</td>
<td>American Dream (.496), Regional Location (.577), Retirement (.471), Travel (.641)</td>
</tr>
<tr>
<td>Culture</td>
<td>Cross-cultural (.885), Open to Others (.456) Fine Arts (.794)</td>
</tr>
<tr>
<td>Religion/</td>
<td>Faith (1.056), Giving Back (.459), Heritage (.542), Keeping Self Pure (.696), Purposeful Life (.511), Putting Others First (.610), Spiritual Beliefs (.980), Tradition (.562)</td>
</tr>
</tbody>
</table>

Table 3. Comparison of health information communication profiles by personality characteristics and cultural factors

<table>
<thead>
<tr>
<th>Measure</th>
<th>Big 5</th>
<th>Cultural Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>FC-H (1)</td>
<td>MC-H (2)</td>
</tr>
<tr>
<td>Extraversion</td>
<td>8.53</td>
<td>158</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>9.64</td>
<td>157</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>9.63</td>
<td>155</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>7.10</td>
<td>159</td>
</tr>
<tr>
<td>Virtue</td>
<td>-0.2</td>
<td>163</td>
</tr>
<tr>
<td>Comfort</td>
<td>0.02</td>
<td>163</td>
</tr>
<tr>
<td>Culture</td>
<td>0.19</td>
<td>163</td>
</tr>
<tr>
<td>Religion and Spirituality</td>
<td>-0.12</td>
<td>163</td>
</tr>
</tbody>
</table>

Discussion and Conclusion

This study demonstrates that there are statistically significant differences in the cultural values and personality traits associated with each HIC profile. These differences can help designers create distinct personas as a foundation for generating consumer health IT solutions that can be tailored to different types of individuals. For example, individuals identifying as OC-H who have strong identification with the “comfort” factor, indicate value placed on travel and regional location. This suggests potential design features such as enabling them to connect with others living in their geographical area and to those with similar health conditions when they are away from home. Moreover, at a general level, this study supports the idea that culturally informed consumer health IT need not only be responsive to culture as defined by race and ethnicity, but also as defined by cultural values that are patient-centered, in the sense that they are generated directly from patients. Future research should continue with an iterative design process with each of these groups to determine how consumer health IT solutions either diverge or converge based upon these distinct personality characteristics and cultural values.

Acknowledgements

This study was sponsored by AHRQ (R03 HS22930-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ.

References

The Discriminative Power of Non-Specific Laboratory Results
Jacob P. VanHouten, MS¹, Christopher J. Fonnesbeck, PhD¹,
Michael E. Matheny, MD, MS, MPH¹, Thomas A. Lasko, MD, PhD¹
¹Vanderbilt University Medical Center, Nashville, Tennessee, USA

Introduction
Physicians generally use all available data when making diagnosis and treatment decisions, but computational approaches to
phenotype identification often limit themselves to a small number of highly-specific, expert-engineered variables when
defining the phenotype of interest. While this approach does provide computational savings, it also limits the sensitivity and
specificity of the phenotype identification.

In this work we distinguish between local and distributed information for an outcome of interest. For example, local
information for diabetes could be an elevated glucose result, the use of metformin, or an ICD code 250, because they are
highly specific for and directly related to the disease. On the other hand, distributed information is non-specific information
distributed throughout the record that is nevertheless related to the presence of the disease, and in aggregate may indicate its
presence or absence. While much research has used local information for phenotype identification¹,², comparatively little has
explored the use of diffuse information³,⁴.

In this work, we estimate the discriminative power of information distributed among non-specific laboratory test results by
applying a standard classification algorithm several different problems. For each problem we also explored the effect of
complexity of data representation, recognizing that greater representation complexity can come at a higher cost in
computational resources and investigator effort.

Materials and Methods
We trained models for eleven different classification problems that had accurate, easily accessible reference standard labels
(Table 1). We selected patient demographic and surgical procedures to represent targets of various degrees In addition, we
trained two models on what we expected to be very difficult conditional problems: did the patient have a kidney transplant,
given that they did receive a kidney or liver transplant, and did the patient have a hip replacement, given that they did receive
a hip or knee replacement?

We used data from Vanderbilt’s Synthetic Derivative, a deidentified mirror of our electronic medical record. We selected the
top 150 most commonly performed laboratory tests, of which we excluded six because they were not laboratory
measurements (dose, patient location, etc.). We limited our study population to the most recent eight years, required that
records have results for at least 10 of the top 144 labs, at least one test for which there were three or more recordings, and no
missing data for sex or race. This left us with a final study population of 357954 records for training and testing.

To evaluate the tradeoffs between prediction accuracy and complexity of computation and data, we tried eight increasingly
complex data representations for each patient record:
1. Binary (whether the test was ever ordered),
2. Counts (the number of orders made for the test),
3. Counts per year for each of the eight years,
4. Cumulative counts by year,
5. Mean (of all results in the eight year span),
6. Quintiles (of all results in the eight year span),
7. Continuous smooth interpolations of result trajectories, and
8. Combination of order counts and result means.

For classifying patients by sex, we performed the analysis with and without prostate specific antigen, which was the only sex-
deterministic test in the suite of labs used.

We selected random forest as our classifier, because typically performs well, even for unbalanced problems; it scales well, it
can learn predictive non-linear combinations of variables, and it can estimate variable importance. We optimized the forest’s
parameters for each task and each representation.

Results
Of the data representations considered, the binary representation performed the worst and the combination of counts and
means performed the best, but means or counts alone were usually close to the best (Figure 1). The easiest problem turned
out to be the detection of kidney transplant, the hardest was distinguishing hip vs. knee replacement. The classifiers using the
continuous smooth interpolations of results could not be run on the entire data set due to memory limits, but analysis using a
subset of data found no improvement over the means representation (data not shown).
The most important variables were not always the same among representations within a specific task. For example, while the presence of a test for thyroid stimulating hormone or urine squamous epithelia was highly discriminative of sex in the binary representation, the actual results of creatinine, hemoglobin and MCHC were the most predictive in the means representation.

Conclusion

We were able to identify patients with specific low-noise findings using only non-specific laboratory tests as input features for random forest classifiers with generally high accuracy. This performance does not require much work to achieve, as models built using lower complexity representations often performed as well as more complex ones. Using result means alone is an efficient way to encode test results. Surprisingly, the timing of orders appears to be less important than whether the orders were placed at all. The extension of order counts to counts binned by year and to year-over-year cumulative counts only performed as well as total counts, not better. This result may be due to the nature of the random forest and how variables are selected for inclusion in each tree. If the sum count of orders is the most information-dense representation, then a random forest classifier would need to select many individual variables from a representation of counts binned by year to encode the same data.

We found that different representations cause the classifier to focus on different variables. It is clinically intuitive that important variables in one representation within a task are not the same as the important variables for a different representation. One possible explanation is that binary and count representations may reflect differences in clinical practice, while laboratory values capture signals related to biological differences between patients with and without the finding of interest. Developing a knowledge repository of these features for generalizable use has great utility.

While we used only random forest classifiers, we expect that our results may extend to other classification algorithms, at least those that are as effective as random forests in extracting complex nonlinear relationships between input variables. Our results suggest that summary measures such as overall means and order counts may be generally efficient data representations that balance performance and cost for common medical classification tasks.

References

Information Extraction from Italian medical reports: first steps towards clinical timelines development

Natalia Viani, MS¹, Valentina Tibollo, MS², Carlo Napolitano, MD², Silvia G. Priori, MD, PhD²³, Riccardo Bellazzi, PhD¹, Cristina Larizza, PhD¹, Lucia Sacchi, PhD¹

¹Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy; ²IRCCS Salvatore Maugeri Foundation, Pavia, Italy; ³Department of Molecular Medicine, University of Pavia, Pavia, Italy

Introduction

Clinical reports include a great amount of valuable information in the form of unstructured text. The need to extract such information to improve medical decisions is well-known. Given that many diseases require continuous monitoring over time, identifying clinically relevant events and eliciting temporal relations among them is crucial. In the field of clinical Natural Language Processing (NLP), many works have been published dealing with unstructured English texts. However, the extent of the related research for other languages is smaller. In addition to the challenges specific to clinical NLP, other language-dependent issues have to be considered. In the case of Italian, the lack of freely available annotated medical corpora, the limited coverage of available dictionaries, and the lack of medical-specific taggers, make it hard to develop techniques to analyze unstructured medical texts. To overcome this problem, several strategies have been proposed. A considerable amount of work for the automatic analysis of both English and non-English texts relies on rule-based approaches. The definition of rules can be enhanced by using domain ontologies, containing information on the main clinical events and their attributes.

In this work we present a rule-based, ontology-driven approach to extract clinical data from unstructured Italian texts. The proposed approach represents a first step towards developing an NLP pipeline that extracts both clinical and temporal information from a set of medical reports, with the final aim of building clinical timelines.

Part of our NLP pipeline has been validated on a set of medical reports provided by the Molecular Cardiology Laboratories of the Fondazione S. Maugeri hospital in Pavia, Italy. Medical reports belong to patients with inherited or acquired heart diseases (e.g. long QT and Brugada syndrome), and cover a time span of 5 years. Many of the data included in reports are currently manually stored in the TRIAD repository (http://triad.fsm.it/triadweb), a database for clinical and genetic variations in the field of inherited arrhythmogenic diseases. It includes data on diagnoses, genetic mutations, cardiac events, performed tests, prescribed treatments and device implants.

Methods

In clinical reports, it is frequent to find occurrences of clinical events that are related to a set of attributes, which characterize the event itself. In the field of Cardiology, for example, ECG test represents an event, while heart rate and rhythm represent some of its related attributes. To explicitly represent this feature, we have defined an ontology, structured into Event and Attribute classes. Attributes can have a numeric (heart rate) or a categorical (rhythm) value. All concepts in the ontology are associated to a specific regular expression, which allows searching for their occurrence in free-form texts.

The NLP pipeline that we implemented is based on Apache UIMA and uses TextPro for standard pre-processing. Negation detection is performed by adapting an implementation of the NegEx algorithm to Italian. Clinical information extraction (IE) is performed in two steps. First, we identify mentions of clinically relevant events, such as diseases, tests and prescriptions. These events are extracted by relying on two external Italian dictionaries: the ICD9-CM, and the FederFarma dictionary for drugs. Each identified event is then linked to its attributes through the developed ontology. Specifically, the ontology modeled in Protegé-OWL is automatically translated into a configuration file used by a specific UIMA Annotator. There are three main reasons why we opted for an ontology-driven approach. First, every time a concept has to be added or modified, a simple ontology update is needed. Moreover, if the analysis of a new set of documents was needed, it would be only necessary to enrich the ontology with new domain-related information. Finally, to analyze a set of documents written in another language, it would be sufficient to modify all the regular expressions.

Given the lack of annotations in the analyzed corpus, to perform the evaluation of the proposed approach, we compared the output of our pipeline to the data stored in the TRIAD database for 4430 reports.
Results

The ontology developed in this work contains 11 events and 60 attributes. To define relevant attributes, the information stored in the TRIAD system was considered. To evaluate our pipeline, we considered a subset of events that are frequently included in Cardiology reports: main diagnoses, drug prescriptions and three diagnostic tests (ECG, Holter ECG and Effort Stress tests). For each report, we compared the values extracted by our pipeline to the corresponding values stored in the TRIAD system. For each extracted event, only those attributes for which an entry was detected in the TRIAD system were considered. Annotations were marked as correct if both the attribute name and its value matched those included in TRIAD. If the attribute extracted by the pipeline was stored with a different value, the annotation was considered not correct. The results obtained in this preliminary evaluation are shown in Table 1. Dealing with drugs was not trivial: while reports often contain separate information on drug names, forms and unit dosages, the TRIAD repository stores only drug names and daily doses. For this reason, the evaluation of extracted prescriptions was performed both on drug names only and on drug names with associated dosages (results marked with * in Table 1).

<table>
<thead>
<tr>
<th>Event Name</th>
<th>Pipeline (a)</th>
<th>TRIAD (b)</th>
<th>Correct Matches (c)</th>
<th>Accuracy (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECG</td>
<td>26669</td>
<td>22546</td>
<td>21352</td>
<td>94.7%</td>
</tr>
<tr>
<td>Holter</td>
<td>26767</td>
<td>21538</td>
<td>19058</td>
<td>88.5%</td>
</tr>
<tr>
<td>Effort Stress</td>
<td>9683</td>
<td>3978</td>
<td>2367</td>
<td>59.5%</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>4202</td>
<td>4077</td>
<td>3607</td>
<td>88.5%</td>
</tr>
<tr>
<td>Prescribed Drug</td>
<td>8720 (8270*)</td>
<td>2436 (4584*)</td>
<td>2186 (2860*)</td>
<td>89.7% (62.4%*)</td>
</tr>
</tbody>
</table>

Discussion

Diagnoses are identified with an accuracy of 88.5%. As regards tests, while the annotation of ECG and Holter tests shows good performances (94.7% and 88.5%, respectively), accuracy drops when considering Effort Stress tests, probably because the results of such tests are described by physicians with articulated sentences. As expected, the evaluation performed only on drug names led to a higher accuracy than the one considering both names and dosages.

The IE procedure that we developed shows promising results. That said, improvements are still needed. First, we plan to extend look-up dictionaries by computing n-grams that denote expressions frequently used in reports. Moreover, improvements on negation detection and misspelled words are needed. By improving the pipeline steps, it would ultimately be possible to build a well-performing system that automatically inserts data into the TRIAD system, saving a lot of manual work and improving data quality. Furthermore, the annotations produced by the system could be eventually used to create a pseudo-gold standard corpus, enabling further research.

References

Evaluation of SMART-on-FHIR I2b2 cell using PCORNET data model

Kavishwar B. Wagholikar, MBBS PhD 1,2, Rahul Jain, MPH CPHIMS, 3,5
Eliel Oliveira, MBS MS, 3,5 Joshua Mandel, MD 1,4, Jeffery Klann, PhD 1,2, Prasad Patil, MS, 5,6 Kenneth D. Mandl, MD MPH 1,4, Shawn N. Murphy, MD PhD 1,2, Thomas Carton, PhD3,5
1 Massachusetts General Hospital, Boston, MA; 2 Harvard Medical School, Boston, MA; 3 Louisiana Public Health Institute, LA; 4 Children's Hospital Boston, Boston, MA; 5 Research Action for Health Network, LA; 6 Persistent Systems, India;

Introduction

Informatics for Integrating Biology and the Bedside (i2b2) is an open source clinical data analytics platform originally sponsored by the National Institutes of Health. (1) It is used at over 100 sites nationally for querying patient data to address clinical questions. I2b2 has been adapted to build multi-institutional networks, and is a central component in the infrastructure for many institutions that have Patient-Centered Outcomes Research Institute (PCORI) awards.

Fast Health Interoperability Resources (FHIR) is a new standard for exchanging healthcare information electronically. It is based on emerging industry approaches, and builds on the lessons from previous healthcare standards. Several projects are underway through standards organizations to facilitate adoption of FHIR including the Argonaut Project, Data Access Framework (DAF) and Clinical Information Modeling Initiative (CIMI).

I2b2 has a SMART-on-FHIR i2b2 cell (SOFI), which serves data from an i2b2 instance in the FHIR format. SOFI provides an alternative, but secure way to retrieve data on per patient basis using OAuth2-SMART specification. (2)

Our objective in the current work is to evaluate SOFI on data in format used by PCORI sites. The goal is to reduce barriers for sharing tools across the PCORI’s patient powered research networks (PPRNs). Our work can significantly reduce the efforts at the PPRNs to develop a platform for running SMART apps, which will facilitate sharing of the apps across the network.

Methods

We imported patient data in CDM format into i2b2 using an i2b2 plugin (3), and adapted the SMART-on-FHIR i2b2 cell (SOFI) to query i2b2 using PCORI ontology. (4) SOFI cell provides demographics, laboratory results and diagnoses for each patient in FHIR format. The deployment was performed at two locations Partners Healthcare (PHC) in Boston, MA and the REACHnet Data Center (RDC) in New Orleans, LA.

For evaluating this infrastructure, we assessed the accuracy and completeness of the FHIR output from SOFI, by performing manual inspections, and by running the FHIR validator available with the specification. We performed spot inspection of 30 instances of FHIR resources (Labs, Diagnosis and Demographics) for random patients at both the locations. We verified the dates, codes and code expansions during the spot inspections, by comparing the FHIR output with the i2b2 data. Additionally, at RDC we traced the data-points into the imported CDM. To evaluate compliance of the SOFI with the SMART OAuth2.0 specification, we developed and executed a dashboard SMART app on both the installations. Finally, we measured the average time required for the platform to respond to request for FHIR resources. (Figure 2)

Results and Future work

The SOFI platform could be installed at both the sites: PHC and RDC. SOFI installation script could be readily run on a CentOS instance at Partners, but needed few modifications for running on Ubuntu instance at RDC’s Openstack node. At PHC we had previously created an i2b2 instance. At RDC’s we created a fresh i2b2 instance by importing CDMs for 300 patients.

The official FHIR validator ran successfully on randomly selected resources at both the sites, and the manual inspections revealed that the FHIR resources were valid and contained accurate data. The SMART app executed successfully at both the sites, demonstrating the security of the installations. SOFI requires an of average 5 seconds for responding to non-cached calls and 0.5 second for cached calls.

We plan to expand the range of FHIR resources served by SOFI on i2b2 instances derived from PCORNET-CDMs.
Furthermore, we are testing portability of apps from SMART gallery across the SOFI installations in PHC and RDN.

Figure 1. Pipeline for installing and evaluating the SMART on FHIR i2b2 cell (SOFI). SOFI provides an installation script that is tailored to virtual machine platform at Amazon. This script automatically pulls the SOFI source code from Github and installs the servlet container, database and webservice on a fresh instance. The pipeline to import data from the CDM format, into i2b2 instance and to install SOFI is available at [http://community.i2b2.org/wiki/display/FCC/FHIR+Connection+Cell](http://community.i2b2.org/wiki/display/FCC/FHIR+Connection+Cell)

![Diagram](image)

Response time in seconds

<table>
<thead>
<tr>
<th>Calls per second</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2. Summarizes the time needed for SOFI to respond to FHIR queries from clients, when the calls are not cached. This incudes the timed needed to retrieve data from the i2b2 instance and translate it into FHIR format. The figure shows that the response times for non-cached calls, plateau at the rate of one call every 5 seconds, for the current server configuration of two cores of 3.3 Ghz with 4 GB RAM (t2.medium amazon ec2 instance). Increasing the configuration (core and RAM) of the server can potentially improve the non-cached response time. For calls that are previously cached the response time averages to 500 milli-seconds.

Acknowledgements: This research is supported by grant under award numbers R00LM011575 from NLM, and R01GM104303 from NIGMS, and CDRN-1306-04864 from the Patient Centered Outcomes Research Institute.

References


Colin G. Walsh, MD MA¹,²,³, S. Trent Rosenbloom, MD MPH¹,²,⁴ Department of ¹Biomedical Informatics; ²Medicine; ³Psychiatry; ⁴Pediatrics Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

In 2011, the American Board of Internal Medicine introduced the Choosing Wisely initiative. Based on empiric research, this initiative delineated healthcare services that should be avoided unless certain “red flag” criteria are present.¹ For example, Choosing Wisely recommends “avoiding imaging for patients with low back pain without ‘red flags’ like cancer or trauma”.² Choosing Wisely has become a widely adopted standard for measuring and promoting healthcare quality. However, Choosing Wisely metrics are currently encoded using International Classification of Diseases (ICD), version 9 codes.³ On October 1, 2015, the Center for Medicare and Medicaid Services CMS) transitioned to the tenth version ICD (ICD10) for billing claims.⁴ Diagnostic coding in ICD10 allows increased code granularity, but resulted in a larger number of available codes. To support transitioning to this larger code set, CMS commissioned the General Equivalence Mappings (GEMs) from ICD9 to ICD10 and Reimbursement mappings from ICD10 to ICD9. Many medical centers also created “one-to-one only” mappings to reflect local billing practices. To date, there are no ICD10 based versions of Choosing Wisely metrics. The purpose of this study is to estimate differences in prevalence assessment for a number of Choosing Wisely Metrics using all available mappings between ICD9 and ICD10.

Methods

This study was conducted using data from the Vanderbilt University Medical Center (VUMC) Research Derivative, a curated mirror of the clinical data warehouse designed to support research. ICD-encoded claims on all adult patients over age 65 at VUMC were collected in two time periods: period A) October 1, 2014 – February 15, 2015, corresponding to when ICD9 was in use; and period B) October 1, 2015 – February 15, 2016, corresponding to when the ICD10 was in use.

This study used Choosing Wisely criteria for: avoiding unnecessary imaging for low back pain; avoiding unnecessary imaging for benign prostatic hyperplasia (BPH); avoiding use of feeding tubes in patients with dementia. All codes from patients in period A were compared against Choosing Wisely ICD9 criteria directly. Codes from period B were mapped to ICD10 codes via the following 4 methods: 1) mapping Choosing Wisely codes to ICD10 using GEMs; 2) mapping Choosing Wisely codes to ICD10 using institutional 1:1 maps; 3) backwards mapping patient codes to ICD9 using reimbursement maps; and, 4) backwards mapping patient codes to ICD9 using institutional 1:1 maps. We calculated numbers of patients having an ICD encoding red flag symptoms among all patients having the underlying clinical condition for each Choosing Wisely metric studied. Cost estimates were based on assumptions in Colla et al.²

Results

In period A, there were 951,187 total ICD9 claims coded on 66,825 individuals. In period B, there were 1,185,198 ICD10 total claims coded on 63,395 individuals. Prevalence of

red flag symptoms varied most using GEMs to map Choosing Wisely codes to ICD10 (i.e., method #1 above). For back pain and BPH imaging, methods 2-4 were closest in estimates to baseline. For feeding tube use, baseline prevalence and cost was 1.2% and $181,658 respectively in Period A. In Period B, estimates ranged from 0.3% prevalence and $66,509 in waste for methods #2-4 up to 1.9% prevalence and $412,106 in possible waste using method #1.

Discussion
This study evaluated how different approaches to converting Choosing Wisely criteria for identifying red flag symptoms from the ICD9-encoded criteria to ICD10. Using Reimbursement Maps (i.e., method #3) led to most similar rates of patients being identified as having red flags compared to baseline, and is likely to be the most generalizable given their public availability. Results also indicate that the method of reconciling metrics based in ICD9 to claims coded in ICD10 may influence estimates of prevalence and cost for Choosing Wisely. This phenomenon was illustrated most by feeding tubes in dementia patients – a relatively infrequent but high cost service if it is implemented against the recommendations of Choosing Wisely.

Strengths of this work include a large dataset of real-world clinical claims and procedure data. Many of the mapping tables used in this work are readily available. Limitations of this study include a lack of a true gold standard in the ICD10 era. There are potential temporal patterns or confounders in clinical operations and workflow as the Choosing Wisely has become well known throughout the country. We are not aware of specific efforts at VUMC that should undermine the expectation that the four mapping methods would yield similar prevalence estimates in the ICD10 era.

Further research in this domain is required. Our analyses should be replicated using data from other centers where billing practices may vary. We speculate that simple mappings of ICD10 codes back to ICD9 or Choosing Wisely ICD9 Criteria to ICD10 is unlikely to achieve parity with what has been established in the era of ICD9. Achieving a highly discriminative algorithm may require a curated gold standard dataset created by professional coders from real-world clinical data.

References

Funding
This work was supported by the CMS Transforming Clinical Practice Grant (PI: Rothman) [CW, STR]
Discovering representational differences between pathway knowledge bases for pathway resource merging
Lucy L. Wang, John H. Gennari, Neil F. Abernethy
Department of Biomedical Informatics and Medical Education
University of Washington, Seattle, WA

Introduction
Biological pathways are powerful tools for analyzing complex biological and pathological processes. There are a large number of pathway knowledge bases, specializing in the biology of different species, functions, and diseases. The contents of these databases are both unique and complementary. In order to perform analyses and draw conclusions over the full breadth of knowledge biologists have generated, we benefit from combining material from multiple pathway databases. Many resources have already begun this process, both by providing resources ostensibly in the same standard (e.g. PathwayCommons), and by allowing for queries and visualization over multiple knowledge bases (e.g. ConsensusPathDB, hiPathDB). However, resource integration remains difficult due to issues of resource quality and differences in knowledge representational choices between resources.

In order to successfully align the contents of these pathway knowledge bases, we must understand how their structure differs and provide a way to translate content between them. Some prior work has been done in describing semantic differences between reaction representations. However, semantics are not the only issue. Several classes of differences exist between pathway resources in entity annotation, entity existence, reaction semantics, and reaction and entity granularity. In this abstract, we focus on distinguishing these variations in content and, more specifically, on detecting inconsistencies in annotation between two popular human pathway resources: HumanCyc and Reactome.

Methods
In the first part of this work, we define a typology of pathway resource differences and provide criteria for identifying differences across five resources: HumanCyc, KEGG, PANTHER pathways, Reactome, and WikiPathways. The following examples use content extracted from HumanCyc (v19.5) and Reactome (v55) in BioPAX format.

Figure 1 illustrates several classes of representational differences in a reaction of the glycolysis pathway, the conversion of phosphoenolpyruvate to pyruvate. The reacting species ADP, phosphoenolpyruvate, ATP, and pyruvate are annotated with cross-referenced ChEBI identifiers, none of which match between the two resources, showing an inconsistency in entity annotation. For example, ATP is annotated as ChEBI:30616 in HumanCyc and ChEBI:15422 in Reactome. HumanCyc also provides one additional participant in this reaction, the H+ ion, an example of a disagreement in entity existence. Next, differences in reaction semantics are seen between the two resources in the representation of reaction direction. In HumanCyc, the entities in green are labeled with BioPAX property “right,” the entities in red are labeled “left,” and the entire reaction as having direction “right-to-left.” In Reactome, the reactants are labeled “left” and the products “right,” with the reaction proceeding “left-to-right.” Lastly, a difference in entity granularity is seen in the catalyzing enzyme pyruvate kinase. Where HumanCyc annotates to two instances of the protein, PKLR and PKM, and their UniProt identifiers, Reactome chooses to represent the enzyme as a complex, the pyruvate kinase tetramer.

Figure 1. A reaction (phosphoenolpyruvate + ADP → pyruvate + ATP) as represented in HumanCyc and Reactome. Cross-referenced identifiers for entities are given above or below their display names. Entities labeled BioPAX “left” are shown in red and those labeled BioPAX “right” are shown in green.
Table 1. Raw counts of proteins and small molecules in HumanCyc and human Reactome, and the corresponding frequency of UniProt/ChEBI annotations for each.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Proteins Subset of proteins w/ UniProt ID (%)</th>
<th>Small molecules Subset of small molecules w/ ChEBI ID (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanCyc</td>
<td>3009 (91.0%)</td>
<td>2467 (65.1%)</td>
</tr>
<tr>
<td>Reactome</td>
<td>19473 (87.0%)</td>
<td>3267 (88.5%)</td>
</tr>
</tbody>
</table>

The second part of this work focuses solely on mismatches of entity annotation. Annotation features (names and identifiers) are the most obvious way of matching entities between resources, so detecting these mismatches are important to performing resource alignment. Entities with the same string name in HumanCyc and Reactome but no shared identifiers are flagged as having potential mismatches in annotation. Samples of these potential inconsistencies are then evaluated to determine whether a true disagreement has occurred and to propose explanations for differences.

Results

The number of unique proteins and small molecules in HumanCyc and human Reactome are given in Table 1. The majority of proteins and molecules in these two resources are annotated with UniProt or ChEBI IDs. String names and identifiers were used to identify potential matches for entities with identifiers. For proteins, we found identifier matches between 2023 HumanCyc entities and 2885 Reactome entities. Of the remaining proteins, we found 208 pairs of complete case-insensitive string name matches between HumanCyc and Reactome that did not share any UniProt identifiers. For small molecules, we found identifier matches between 387 HumanCyc entities and 701 Reactome entities. Of those small molecules that did not have identifier matches, 1869 pairs of string name matches between the two resources were found, indicating potential identifier annotation mismatches.

Upon inspection of these annotation inconsistencies, we found several explanations for mismatches. For proteins, annotation with UniProt secondary accession numbers explained some of these differences. Secondary accession identifiers redirect to primary identifiers, and refer to synonymous entities, e.g., UniProt:Q96NC3 redirects to UniProt:Q6NXT4 (Zinc Transporter 6). For small molecules, the vast majority of ChEBI identifier mismatches result from annotation with conjugate acid/base pairs from ChEBI. In Figure 1, all four reaction participants show annotation to conjugate pairs in ChEBI, i.e., ChEBI:32816 (pyruvic acid) is a conjugate acid of ChEBI:15361 (pyruvate) and so on. The ChEBI identifiers can also point to tautomers, e.g., ChEBI:16828 (L-tryptophan) and ChEBI:57912 (L-tryptophan zwitterion); or groups instead of specific entities, e.g., “glyoxylic acid” references ChEBI:16891 (glyoxylic acid) and ChEBI:35179 (2-oxo monocarboxylic acid), and the former is an instance of the latter, which refers to a group of molecules.

Using string names to identify potential matches has its limitations. Some proteins in the two resources, for example, may have the same name but refer to different entities, e.g., ELK1 is an abbreviation for an ETS domain-containing protein or an Ether-a-go-go-like potassium channel. Therefore, although two proteins match on string name between the resources, there is a valid reason why they refer to different entities.

Discussion

Identifying and resolving differences between biological pathway resources is challenging. However, it is a necessary prerequisite to faithfully interpreting, integrating, and using the content of multiple resources. We described a typology of differences between pathway resources, and then quantified and explained annotation mismatches between Reactome and HumanCyc. Exploring entities that matched on string names but not identifiers offered some insight into why some of these annotation mismatches may occur. Our results suggest that computational tools can be built to aid alignment of these differences in representation. Through alignment and integration, we can better leverage our knowledge of biological pathways, and build next-generation tools for pathway analysis and drug targeting.

Acknowledgements

This research was supported by the National Library Of Medicine Training Grant T15LM007442.

References


155
Modeling Flowsheet Data for Secondary Use

Bonnie L. Westra, PhD, RN, FAAN, FACMI1,2, Beverly A. Christie, DNP, RN3, Matthew D. Byrne, PhD2, RN4, Anne LaFlamme, DNP, RN5, Grace Gao, DNP, RN4; Steven G. Johnson, PhD-C2, Jung In Park, PhD-C1, RN, Lisiane Pruinelli, PhD-C, MSN, RN4, Piper A. Ranallo, PhD6, Suzan G. Sherman, PhD, RN3, Connie W. Delaney, PhD, RN, FAAN, FACMI1,2, Stuart Speedie, PhD, FACMI2
1University of Minnesota, School of Nursing, Minneapolis, MN; 2University of Minnesota, Institute for Health Informatics, Minneapolis, MN; 3Fairview Health Services, Minneapolis, MN; 4Optum Health, Bloomington, MN; 5University of Minnesota Health, Minneapolis, MN; 6Six Aims for Behavioral Health, Minneapolis, MN

Introduction

Big data analytics is emerging as a critical way to address precision medicine, personalized health care, and research. Interprofessional assessments and interventions represent unique perspectives for health promotion and illness prevention, responses to diseases, self-care, population health, family and community health, and behavioral and social determinants of health. This data is frequently captured in a document type called flowsheets; however, interprofessional data is seldom integrated into clinical data repositories in a useful way that supports secondary use, such as research. Flowsheets are templated documentation forms similar to spreadsheets and include: the names of the observations are located in the first column and documentation by date and time in subsequent columns. In some electronic health records (EHRs), a screen view of flowsheet data is called a template which can be unit or discipline specific; contain groups of related assessments and interventions; and individual items (flowsheet measures) for actual documentation of care. There are beginning efforts to model flowsheet data for comparison across settings; however, there are numerous challenges to making flowsheet data useful for research. Challenges include the large volume of data which can slow the extract, transfer, and load process; duplication of semantically comparable flowsheet measures; institutions’ preferences and customizations; software upgrades; and, the lack of standard coding for the types of data captured. Thus, one post-implementation solution is to create information models that include concepts about a specific topic e.g. pain, and organize concepts in a hierarchical manner. These information models then can be used to map semantically comparable flowsheet measures to concepts and subsequently link the concepts to standardized terminologies. The purpose of this study was to derive information models from flowsheet data using a data-driven information modeling process.

Methods

An iterative data-driven consensus based method was used to create information models that hierarchically organize concepts about a specific topic and map flowsheet observations to the concepts. This method is similar to that used by Harris et al. when creating a pressure ulcer model from six EHRs. After IRB approval, a random sample of 199,665 de-identified encounters representing 66,660 patients who received care in a Midwest health system between October 20, 2010 and December 27, 2013 was transferred to a secure data shelter. The flowsheet data represents 34% of all of the observations in the clinical data repository (CDR).

Data in an EHR is designed from a clinician’s viewpoint and needs to be modeled differently for secondary use, such as research. Actual de-identified patient flowsheet data documented in the EHR were analyzed. Two master files were created in Excel to demonstrate both the documentation context and counts of unique flowsheet measures distributed across the entire EHR regardless of the context where documentation occurred. The data were extracted from a CDR to create these master files. The documentation context included the names and unique identification (ID) numbers for templates, groups of similar observations, and the flowsheet measures. The file of just the unique flowsheet measures included names and IDs aggregated across the EHR along with their value type (i.e. date, numeric, choice list, etc.), the responses documented, and the frequency of use.

Multiple steps were used to create the information models: identifying topics and related concepts, searching master files for relevant flowsheet measures, and creating a new Excel file linking flowsheet measures to flowsheet measures to Concepts and organizing the Concepts into Classes (groups of observations). Selection of topics was based on five quality measures prioritized by one of the health system research partners, documentation for five high volume physiologic systems, and four behavioral health national priorities. Relevant concepts were identified from knowledge sources, clinical expertise, and discovery when searching the flowsheet data. A manual process was used
to search the master files for flowsheet measure representing concepts and concept synonyms; flowsheet measures were copied and pasted into a new Excel file linking the measures to the information model Concepts. The Concepts and linked flowsheet measures subsequently were organized logically into Classes in the information model. This process is similar to that used for creating a pressure ulcer information model in the Federal Health Information Model (www.fhims.org/press_ ulcer.html). A second investigator reviewed the findings for each information model by affirming the mapping of flowsheet measures to Concepts and in addition, searched for any flowsheet measures that may have been missed. Results were presented to the entire team during weekly conference calls to obtain consensus on the information models and mappings. After all models were created, they were further reviewed by two of the investigators (BW & SS) and any changes were presented to the research team for consensus.

Results

There were 14 information models created and validated by the research team using a consensus process. These models include the following topics: Cardiovascular System; Gastrointestinal System; Genitourinary System (including CAUTI); Neurumusculoskeletal System; Peripheral Neurovascular System (including VTE); Respiratory System; Pressure Ulcers; Pain; Falls (Safety); Vital Signs, Height/ Weight; Aggression and Interpersonal Violence; Psychiatric Mental Status Exam; Suicide and Self Harm; and Substance Abuse. An example of one information model for Genitourinary System (including CAUTI) is shown in Figure 1.

Conclusion

In this study an iterative data-driven process was used to develop and validate 14 information models. While the 14 models were developed using consistent rules, there is no overarching information model that links between the models. The next steps include validation of the information models by six additional large organizations to increase the generalizability. Subsequently LOINC and SNOMED-CT will be used to link the observations to national terminologies. These information models build on work by Harris and colleagues and extends the LOINC coding of assessments captured in flowsheet data initiated by Matney and Settergren’s workgroup, which is part of the Nursing Knowledge Big Data Science Conference.

References

Barriers and Facilitators to Using Electronic Health Records for Referrals between Primary and Specialty Care Clinics

Justina Wu, MPH1, Laura G. Militello, MA2, Mindy E. Flanagan, PhD3, Barry C. Barker3, Shakaib Rehman, MD4,5, Brian W. Porter3, Jasma M. Adams3, April W. Savoy, PhD3,6,7, Alissa L. Russ, PhD3,6,8, Michael Weiner, MD, MPH3,6,7

1Center for Innovation to Implementation, VA Palo Alto Health Care System, Menlo Park, CA; 2Applied Decision Science, LLC, Dayton, OH; 3Center for Health Information and Communication, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service CIN 13-416, Richard L. Roudebush VA Medical Center, Indianapolis, IN; 4Phoenix VA Health Care System, Phoenix, AZ; 5University of Arizona, Phoenix, AZ; 6Regenstrief Institute, Inc., Indianapolis, IN; 7Indiana University, Indianapolis, IN; 8College of Pharmacy, Purdue University, West Lafayette, IN

Introduction

Electronic health records (EHRs) offer to streamline referral processes by facilitating information exchange between referring and consulting providers, allowing them to document and communicate findings. In practice, poor coordination and communication gaps continue to challenge EHRs’ effectiveness, resulting in delays in care. In the U.S. Department of Veterans Affairs (VA) health system, among a sample of referrals in 2013, consultants evaluated Veterans within 90 days in only 28 (19%) of 150 referrals reviewed. The objective of this study was to identify barriers and facilitators to effective referrals and consultations between primary and specialty care clinics, and to suggest how EHRs can make this process more efficient.

Methods

Our interdisciplinary study team included experts in human factors, informatics and clinical medicine. Eight individuals conducted semi-structured interviews and ethnographic observations of referrers, consultants and other key roles involved in managing referrals from primary to specialty care at two large, geographically dispersed VA medical centers serving urban and rural patients. We sampled participants by inviting clinicians, nurses, triage teams, and support staff involved in the consultation process to participate. We recruited in-person and via email from primary care and six subspecialties selected to obtain a balance of high-volume medical and surgical services (Table). Four researchers analyzed interview data using inductive qualitative methods, synthesized emergent themes related to barriers and facilitators, created a codebook, and resolved coding discrepancies by consensus. They also reviewed observation field notes to create vignettes and workflow diagrams of the consultation process, providing context for interpreting interview data. VA Central Institutional Review Board approved this research.

Results

We conducted 61 interviews and 38 observations with 53 primary care and 46 specialty care clinicians, nurses and support staff (Table). Participants identified barriers and facilitators of EHR-based communication, additional modes of information exchange, and factors outside the EHR that influence the quality of the referral process. EHR-based request forms are customized by each specialty service. Referrers and consultants cited common frustrations with the ability of the forms to facilitate communication about a referral. Referrers noted great variability in quality of forms. Some forms “ask a bazillion questions” while others “will tell you what the prerequisites are and some of them won’t.” Forms that facilitated referrals were “brief,” asked for “just the most basic information,” and allowed users to enter text in form fields. Consultants, however, noted receiving incomplete referrals, including forms that did not specify a clear medical question. If a referral was incomplete or should be handled by another service, a consultant may deny the request through the EHR. Referrers, however, complained that, “Rather than sending you a note saying, ‘Hey, we need this test done first,’ [consultants] cancel it” without further explanation.

To avoid these communication breakdowns, some providers used additional EHR features to communicate about a consultation, but noted that this non-standard practice can lead to lapses in communication: "I can write a note on the consult, and the person who wrote the consult should see it and respond. Should. I know from habit certain people do and certain people don’t."
Because information exchange via the EHR was insufficient, providers also used phone, email, IM or in-person methods, often trying multiple modes due to variability in communication preferences. Providers noted that face-to-face contact with other services facilitated referrals: “I’ve been successful if I’ve gone to visit them physically, if I took the elevator ride and went knocking on somebody's door and said here's what I need.”

Referrers and consultants reported disagreement about responsibility for tasks such as ordering tests and communicating results to patients. One referrer said, “[Consultants] should just…order whatever test that they are gonna be the ones that are gonna be following up on it.” Understaffing and VA-wide mandates to see a patient within a certain number of days also put pressure on consultants to deny referrals or transfer patients to other services for care, leading to further delays. Finally, referral requests to non-VA specialists create their own challenges. Lack of EHR interoperability between VA and outside health facilities force providers to use fax and scans, and even rely on patients to carry their own information.

Discussion

Although EHR-based request forms are designed to improve quality and speed of referrals, participants indicated that their automated and rigid nature fostered rather than eliminated inefficiencies. Improvements to referral processes should include a fresh examination of how to design and use EHRs to enhance communication between referrers and consultants, while optimizing completeness of information. In addition, EHR systems cannot completely replace interpersonal communication regarding consultations. Instead, services should establish agreements about roles and responsibilities within the referral, consultation and follow-up processes. Problematic referrals should prompt direct discussions between consultants and referrers, potentially decreasing cancellations and wait times for patients. EHRs should include more guidance for referrers, as well as pre-populated diagnostic test results or other pertinent details.

Our findings have relevance for both VA and non-VA care. Many problems identified relate to lack of EHR interoperability and can be seen in health systems nationwide. Significant wait times for care outside the VA may also be related to EHR-based processes. In a 2014 national study, the average cumulative wait time to see a physician in five specialties was 18.5 days, with a range of one to 256 days. Our study has limitations. Participants may not reflect all specialties in medicine, and the targeted medical centers may not reflect all medical centers, but we have no specific evidence that our findings are skewed.

Principles of effective consultations have been known for decades, but our findings suggest that opportunities still remain for improving referral communication between requesting and consulting services.

References

Policies and Procedures on Governance of Data Use to Support Health Information Exchange in Low- and Middle-Income Countries

Lauren A. Wu, MHS1,2, Theresa Cullen, MD, MS1
1Regenstrief Institute Global Health Informatics Group, Indianapolis, IN; 2Office of the National Coordinator for Health Information Technology, Washington, DC

Abstract

Low and middle-income countries (LMIC) are increasingly contemplating electronic health data sharing architectures using the health information exchange (HIE) approach. Successful large-scale interoperability depends on effective policies and procedures for data sharing. We undertake a review of policies and procedures for HIE in LMIC to determine the current state of and identify a high-level potential pathway to successful development and implementation of policies and procedures focused on data use and governance.

Introduction

Clinical care, community surveillance, and program evaluation activities are enhanced by a milieu where health data is easily shared. Health information exchanges (HIE) are a common architectural approach for electronic health data sharing, but their success is dependent upon the development and implementation of data governance at a country or institutional level. As low- and middle-income countries (LMIC) move towards increasing HIE and interoperability, developing a logical framework to resolve these sociotechnical ambiguities for a given environment becomes more critical.1,2 We undertook a global review of policies (e.g., laws, national strategies, legal agreements) and procedures (e.g., implementation guidance, standards) (P&P) for health information exchange in LMIC, focusing on data use and governance. Our goals were to add to the existing knowledge and identify key focal areas for a potential framework that country decision-makers should consider when developing effective P&P for HIE.

Methods

We performed a mixed-method review of P&P for HIE in LMIC as defined by the World Bank from February-March 2016. HIE was defined as either the verb of electronically exchanging health-related data or the noun in the form of organizations dedicated to the secure exchange of health-related data (http://www.himss.org/library/health-information-exchange).

For the non-systematic literature review, we reviewed the published literature indexed in PUBMED and MEDLINE using keywords “low- and middle income country,” “health information exchange,” “interoperability,” “big data,” “policy,” and “governance.” We also reviewed articles referenced by the published literature as well as the gray literature including countries’ eHealth strategies, legislation, and other P&P relating to HIE.

For the semi-structured interviews, we contacted individuals working in LMIC through the OpenHIE community (http://ohie.org). The OpenHIE community improves the health of the underserved through open and collaborative development and support of country-driven, large scale health information sharing architectures. Experts varied in expertise but generally have been involved in the development and implementation of country eHealth strategies or direct implementation of HIE systems within countries. Experts were invited to individual or small group (by country or geographic area) hour-long semi-structured phone interviews regarding P&P for HIE. When phone interviews were not possible, experts provided written information via e-mail. This review was qualitative in nature and thus the results presented represent the general themes and overall findings. It does not represent a quantitative perspective.

Results

Literature Review and Interviews. We reviewed 32 articles (9% from LIC, 22% from MIC, 38% from a mix of MIC and LIC, 31% generally applicable to all countries) and 125 documents from the gray literature (72 P&P documents; 20 reports, guidance, and toolkits; and 33 eHealth strategies). Twenty-nine LMIC interviews (or email interactions) were performed (21% from LIC, 38% from MIC, 35% from a mix).

Use Cases for HIE. We gathered information on 28 LMIC and one regional network in Asia. The use cases and value proposition for HIE vary and appear to reflect country and funder priorities. For example, in some African and Caribbean LMIC, HIE focuses on infectious diseases, often due to program monitoring and evaluation (M&E) requirements of donors. Other LMIC, such as in Asia and South America, are developing enterprise-wide HIE with

160
a goal of universal health coverage. Most LMIC are prioritizing data exchange within the country rather than cross-border data exchange. Only one regional network in Asia was identified as working on two fronts—exchange within countries and cross-border exchange through a regional HIE “lab” structure that has a foundational commitment to the use of common standards and formats.

**eHealth Strategy and Data Stewardship/Ownership.** Countries’ eHealth strategies and HIE approaches generally address high-level governance and partnerships; enterprise architecture; patient identification; standards for interoperability; and privacy and security (P&S). In most countries, the data collector or the place where data is physically located is seen as the “owner” or steward of the data. Generally, this has been the Ministry of Health (MoH) in the public sector; where there is a strong private sector, regional or local hospitals and clinics can also “own” or steward the data. However, there is an increasing shift toward discussing the patient as the true owner of their health data. Despite the shifting mindset, few LMIC offer individuals access to their electronic health data.

**Governance.** While all countries’ eHealth strategies featured some governance structures, their implementation and maturity vary. In countries focusing on enterprise-wide architecture and universal health coverage such as in South America and Asia, governance tends to take priority and is implemented early. Countries with a strong centralized government and/or MoH have been able to develop and implement governance from the “top-down,” but enforcement remains an issue. Where there are health issue silos and a strong dependence on donor and private sector funding, governance may not exist nationally and may vary in structure by the health issue.

**Data Use, Exchange, and Patient Identification.** A number of countries have concerns about how data will be reused; this has manifested through policies that require individual-level data remain at the site where the data was collected to one-way data flow to the national/MoH level but not back down to the point of care for real-time care decision makers. Data are frequently reported to the MoH at aggregate levels, supporting M&E reports for donors. However, most LMIC envision HIE supporting individual level health exchange. As expected, patient identification was identified as an important concept for matching records.

**Legal Tools for Data Exchange, P&S.** Where the MoH is the main player and policy setter, there normally exist data use and exchange guidance promulgated from the MoH. However, only one LMIC informant in the review indicated a legal data use and exchange agreement was in place. While many LMIC are investing in data infrastructure and standards, the majority indicated challenges in establishing business agreements for data exchange. Some LMIC have laws or policies in place regarding P&S of health information. There appears to be a dearth of overarching legal frameworks for P&S, though procedures have usually been established through individual projects and implementations.

**Discussion**

These findings provide a snapshot of HIE efforts in LMIC across the globe, illustrating significant variability in the completeness of and success implementing policies and procedures that support in-country HIE and governance of data use. Great attention is being paid to HIE; however LMIC have experienced challenges moving from strategy to implementation. Variation exists around the establishment and implementation of P&P for use cases, governance models, data stewardship, patient identification, composition of legal agreements, and the role of P&S. This review identified a major high-level barrier is the lack of an agreed-upon policy or procedure in the aforementioned areas at the country level directing HIE decision-making and implementation, indicating a need for tools to help countries establish and implement such policies and procedures.

Technology solutions are available for HIE but the implementation of business processes and legal tools, such as data use agreements, have lagged behind the technology. Based on these findings, we suggest that 1) countries and regions should ensure there is a consistent governance model for HIE and interoperability, and 2) development of a global framework for the key policy and procedure decisions supporting HIE and interoperability can accelerate the actualization of data sharing, leading to improved data use and better health outcomes. A global framework for HIE policy and procedures will also need to examine the costs and feasibility of policy and procedure decisions.

**References**

Probabilistic Population-level Modeling of Disease Event Timelines

Stephen T. Wu, PhD\textsuperscript{1,2}, Yanshan Wang, PhD\textsuperscript{2}, Sunghwan Sohn, PhD\textsuperscript{2}, Chung-il Wi, MD\textsuperscript{2}, Elizabeth Krusemark\textsuperscript{2}, Hongfang Liu, PhD\textsuperscript{2}, Young J. Juhn, MD\textsuperscript{2}
\textsuperscript{1}Oregon Health & Science University, Portland, OR; \textsuperscript{2}Mayo Clinic, Rochester, MN

Introduction
The relative timing of a patient's clinical event can be crucial in diagnosing and treating chronic diseases. But timely identification of disease events is a challenging task, even for medical professionals: Molis et al. reported that 65% of their asthmatic population had a delayed diagnosis of asthma with the median delay of 3.3 years (1). In this work, we take a data-driven approach, probabilistically modeling the temporal relationships between clinical events in patient timelines. To aid generalization across patients, we build our models on population-level data.

As part of our previous work on patient-level temporal aggregation (2), we utilized kernel density estimation (KDE) to estimate relationships between observed natural language processing (NLP)-derived features (e.g., nighttime cough, bronchodilator use) and asthma status. More specifically, we defined probability density function (PDF) embodying temporal relationships between an observed asthma-indicator-feature and an observed asthma status (where observations are themselves events extracted from the text of clinical notes in the EHR).

In this work, we present a generalization of our previous temporal modeling framework (2) to arbitrary sequential events (e.g., any time-stamped symptom, test, or disease status). Also, we report the implementation of this probabilistic, population-level model in a corresponding to-be-released open-source software package, adept (Aggregation of Disease Evidence for Patient Timelines), and illustrate its capabilities on the use case of pediatric asthma progression (i.e., disease status progresses as: no asthma, positive asthma, remission, then relapse).

Structured clinical data has been visualized effectively by systems such as LifeLines (3) or TimeLine (4), these tools primarily enable interaction with the data. While our work also deals with categorical variables expressed temporally, our work enables automatic computation utilizing the data. Namely, probabilistically modeling these temporally situated events lays the groundwork for more accurately reasoning at the patient level.

Methods
First, we extend our previous framework (2) to allow for multiple sequential events in a patient timeline, beginning with the Index Date Temporal Density. We defined probability density functions (PDFs) connecting features to asthma statuses as \( p(t) = P(\text{index\_date} = t | f(0) = 1) \). The index\_date is defined as the first date on which a patient should have been considered to have asthma, by manual record review. Binary features (such as a "wheezing" symptom) as discovered by NLP techniques in a clinical note are denoted as \( f(\tau) \), where \( \tau \) is a relative indicator of time. Below, \( t \) is the time elapsed since feature \( f \) was observed at time \( \tau = 0 \).

We could rewrite this equation to straightforwardly extend the model to multiple event types:

\[
\text{Event Start Temporal Density} \quad p(t) = P(\text{start\_date}(e_b) = t | e_a(0))
\]

For asthma, \( e_b \) could be \{no, asthma, remission, relapse\}, and \( \text{start\_date}(\cdot) \) would choose the first date with documents attested in features taking a particular value.

More generally, we model the probabilistic temporal relationship between two event timelines, \( e_a \) and \( e_b \), which are each an array of time-stamped event values. Since event timestamps in \( e_a \) and \( e_b \) may not coincide, we first do an extrapolating step, inferring event values at a specified set of dates/times. Subsequently, we do a differencing step, in which the temporal gap between event timelines are calculated as samples.

For example, consider the timelines \( e_a = (\text{broncho}, t_0), (\text{broncho}, t_2), (\text{nobroncho}, t_3), (\text{broncho}, t_4) \) and \( e_b = (\text{asthma}, t_1), (\text{remission}, t_2), (\text{relapse}, t_4) \). These show the correlation of bronchodilator use with asthma progression. The \( \text{start\_date}(\cdot) \) function is a no-op extrapolating step. The samples used to estimate the temporal densities for remission are the output of the differencing step: \( t_3 - t_0, t_3 - t_2, 0, t_3 - t_4 \). Further generalizations of the models, which we elide here for space, employ other extrapolating and differencing strategies and allow us to handle absent symptoms differently than missing data.
We implemented the estimation of temporal densities (and the aggregation methods from previous work) in a new Python package called \textit{adept}, which will be released to the open-source community as future work.

\textbf{Results}

Preliminary analyses were carried out on a small subset (n=35) of the Mayo Clinic Sick Child Cohort used in previous studies. Each member of this subset had been identified by manual chart review as having asthma; in addition, 17 of these progressed to remission; 10 of those 17 patients then progressed to relapse. We excluded 2 cases where the patient experienced a second remission after relapse.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{temporal_densities.png}
\caption{Several temporal densities. In each plot, time $t=0$ corresponds to the time at which a FEATURE (e.g., \textsc{asthma\_medication} in the legend) is observed to be present. Each line is a different FEATURE $e_{jt}$; features shown here were the most common of those designed to detect remission and relapse. Each plot shows a different asthma status as $e_{jt}$. The x-axis is relative time in days, while the y-axis is probability mass (area should sum to 1.0).}
\end{figure}

Figure 1 shows temporal relationships between features (each line) and 3 stages of asthma progression (each plot). The features plotted here were designed by medical experts to detect asthma remission and relapse. The top plot shows significant probability mass to the left of $t=0$ for all features. This indicates that significant symptoms and tests were observed subsequent to the initial diagnosis of asthma. The middle plot shows significant probability mass at about 1000 days after a feature is observed; thus, remissions typically occurred about 3 years after (corroborating our clinical criteria for remission) asthma symptoms abated. The bottom plot does not have a consistent trend, showing that the relationship between this feature set and relapse may not be meaningful.

\textbf{Conclusion}

We have probabilistically modeled the relationships between multiple, temporally-situated chronic disease events at a population level, and reported its implementation in \textit{adept}. In future work, we will utilize a larger data set, explore multivariate modeling, and measure the effect of these models on evidence aggregation.

\textbf{References}

Transferring Homegrown Medication Knowledge Assets to a Commercial EHR: Overview of the Approach and Lessons Learned

Eileen Yoshida, RPh, MBA, Karen Bavuso, RN, MSN, Shirley Xiang Fei, RPh, PharmD, Ronelle Stevens, RPh, PharmD, Ana Delgado, RPh, PharmD, , Irina Kofman, RN, Saverio Maviglia, MD, Roberto Rocha, MD, PhD

Introduction
In 2012, our institution decided to replace our internally developed clinical systems with an integrated commercial electronic health record (EHR) system, supported by a commercial medication knowledge base vendor. We will describe our experience in migrating internally curated medication clinical decision support (CDS) assets to a commercial EHR that primarily utilizes medication CDS assets from a commercial knowledge vendor.

Methods
Partners Healthcare is a Boston-based integrated healthcare delivery network. For many years, the different facilities have been using different clinical systems within their organizations, duplicating efforts to develop and maintain knowledge assets and functionality, and also propagating inconsistency in practice. Over the years, several proprietary medication knowledge bases and CDS were shared among some of our institutions. These activities have been described and evaluated in detail. We performed a gap analysis for five of our proprietary knowledge bases to identify the key differences in functionality between current and future state. A structured form guided the analysis. Representatives from the EHR vendor participated. Potential future state implementation options were identified. The executive summary of the gap analysis and proposed options were presented to the governance committee responsible for CDS within the new EHR. The CDS Committee includes approximately 40 clinician members from across the enterprise, representing a variety of clinical roles, including our Quality and Safety group.

Results
Please see summary table of decisions made for the five knowledge bases (Table 1). There were differences in both content and functionality that needed to be reconciled. The most prominent functionality differences were (i) the lack of tiered user interfaces for different severity levels in the vendor system (all alerts are interruptive, and hard stops are not supported), (ii) the presentation of all medication warnings generated in an order session altogether at the time of signing instead of individually at the time the warning is generated by an order, and (iii) the inability to accommodate non-commercially available medications (such as investigational drugs and compounded agents).

Discussion
Kuperman et al proposed customizing commercial medication knowledge bases for local use and also as a means to share the burden of knowledge development and maintenance. Our presentation will summarize our experience in the context of these recommendations and feedback from users. We were able to customize the commercial medication knowledge base to filter out specific interventions, and to adjust severity levels using both knowledge base and EHR tools. In some cases, we successfully added new rules to the vendor knowledge base, or created new rules to maintain existing content. Varying levels of effort were required. In other cases, we were not able to replicate original content and functionality due to tool limitations. It was sometimes difficult to ascertain if a rule from our homegrown systems existed in the vendor knowledge base since the rules were expressed in different formats and terminologies. Better tools for browsing and comparing the knowledge and standardized rule architecture and terminologies would be helpful. Content that was readily accepted in our homegrown systems was not always accepted in new system, given functionality differences. While progress has been made, our work demonstrates that opportunities for future development still exist. Commercial medication knowledge base solutions may be a good starting point, but organizations like ours invariably need to be able to extend the knowledge and customize using both knowledge base and EHR tools.
Table 1: Summary of Results

| Duplicate Therapy: 34 classes based on medication vendor classes, augmented with non-commercial medication concepts (investigational drugs and compounded medications); customizable messages; alerts fired interruptively and individually as each order entered. | Vendor EHR recommends enabling all available 861 duplicate therapy classes from medication vendor using 2 different severity levels; alert fires in the Medication Warning screen during order signing (with all other medication warnings generated in the ordering session). | Implemented 34 duplicate therapy classes to match initial state using one severity level; non-commercial medications were included; after 7 months of experience, are starting to turn off these alerts in certain phases of care; small work effort to mimic the current state content, using EHR tools. |
| Drug pregnancy alerts (n=1800): Warnings fire for patients who are pregnant (as inferred by lab values), or of child bearing age, for medications with FDA drug pregnancy category D, X at time of ordering; in addition, implemented a custom category of very high risk medications; 3 levels of severity displayed in tiered presentation. | Medication vendor knowledge base contains FDA drug pregnancy category information for all medications; vendor EHR recommends displaying warning in the Medication Warning screen for categories D, X only; warnings appear together with all other medication warnings at the time of signing new orders, if “Pregnant flag” has been manually set by a user (lab values and age/gender are not considered). | Implemented drug pregnancy warnings for categories D, X; alerts fire based on manually set “pregnancy flag”; to accommodate the custom category of high risk medications, a different type of alert was used; small work effort to configure medication warnings to fire for only categories D, X; medium work effort to develop additional alerts for very high risk medications; using vendor EHR provided tools. |
| Drug-drug interactions (n=8,500): 3 levels of severity were displayed in tiered presentation to ordering providers at the time of ordering; locally vetted knowledge base; includes medication concepts not available in commercial knowledge base. | Medication vendor knowledge base contains ~500,000 DDIs categorized into 4 severity levels; Vendor EHR recommends customizing medication vendor content using medication vendor tools; alerts display to user together with all other medication warnings in the Medication Warning screen at the time of signing (but no ability to implement hard stop or non-interruptive warnings). | Implemented only those DDIs available in initial state; non-commercial medications included; unable to implement all DDIs (~36% not available in vendor knowledge base); submitted to vendor knowledge base for consideration, not all accepted; after 7 months, based on feedback from users citing alert fatigue, filtered out DDIs which were non-interruptive in initial state; large work effort using medication vendor tools. |
| Dose modification for elderly and renally impaired patients (n=350 and ~280, respectively): Modified dose list with default dose and modified frequency presents based on patient’s age or renal function, respectively; interruptive alert will also appear if substitute medication is recommended; locally vetted messages; trumping rules can be set between 2 knowledge bases so only one set of rules fire. | Vendor EHR recommends using medication vendor dose range checking module for both; also recommends geriatric precaution module for elderly; former provides inline warning messages when ordered dose falls outside set of “min/max” range; non-interruptive if dosing is within range, but interruptive if dosing exceeds min/max; latter provides interruptive alert for medications contraindicated or strongly recommended to be avoided in elderly. | Implemented dose range checking; implemented geriatric precautions for medications in initial state; to accommodate substitute medication functionality we implemented a different type of rule; small work effort to configure vendor knowledge bases; large work effort to develop additional alerts for substitute medication functionality. |

References

Data-driven Clinical Pathway Learning and Outcomes Prediction for Chronic Kidney Disease

Yiye Zhang¹, Rema Padman²
¹Department of Health Policy and Research, Weill Cornell Medicine, New York, NY, USA; ²The H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA

Introduction
The best available medical evidence for clinical management of complex, multi-morbid patients is conveyed to frontline healthcare providers through clinical practice guidelines (CPGs) and CPG-derived clinical pathways. Currently, most clinical pathways adapt evidence-based recommendations from CPGs to local practice workflows, and are used by more than 80% of all US hospitals. Clinical pathways aim to reduce variations in treatments, support clinical decision making when faced with multiple or ambiguous care options, improve quality of care, and control costs. However, there are significant challenges in using them at the point of decision making and tracking the continuously evolving impact of interventions on disease progression. CPG-driven clinical pathways are known to suffer from drawbacks such as lack of physician agreement, and most importantly lack of consistency on the strength of evidence underlying them.

An exemplar is Chronic Kidney Disease (CKD), a complex chronic condition that affects 1 in 10 US adults, where patients gradually lose their kidneys’ functions in 5 stages. Most CKD patients suffer from comorbidities such as hypertension and diabetes, resulting in higher medical expenses than many other health conditions. In addition, CPGs for CKD are often consensus-based, due to the lack of large and high-quality clinical trials. Early identification and treatment of individuals at risk as they progress through the 5 stages of CKD can potentially mitigate adverse events such as episodes of acute kidney injury (AKI), hospitalizations, and death. This research investigates data-driven clinical pathway development by analyzing CKD patients’ demographic and detailed treatment data captured in electronic health records (EHR) as part of routine care delivery. We propose a clinical pathway learning algorithm that models 5 types of clinical information: encounter type, diagnosis, procedure, medication, laboratory records, and time, to deduce data-driven clinical pathways and provide predictions for patients’ future health conditions. Leveraging a rich and unique clinical dataset extracted from the EHR of a community nephrology practice, this study learns clinical pathways from data by modeling the longitudinal, multi-dimensional treatment process as a hidden Markov model (HMM).

Methods

Existing clinical pathway learning models tend to produce spaghetti-like structures that are too complex to interpret, make unrealistic assumptions for disease management, or are challenged by the number of clinical activities that can be analyzed and discovered. In this research, each patient’s clinical history is organized as a chronologically ordered pathway of visits to providers and hospitals. Patients’ pathways diverge as their health conditions and corresponding treatments evolve in different ways. Therefore, we use longest common subsequence (LCS) distance metric to measure pathway similarity regarding clinical factors and treatment frequencies. Further, hierarchical clustering is used to cluster patients into subgroups based on the similarity of their pathways, measured with LCS distance. Within each subgroup, each patient pathway is modeled as a HMM, comprising of a sequence of observations and a corresponding sequence of hidden states. The sequence of hidden states is a time-homogeneous
\( k \)-th order Markov chain\(^4\), where \( k \) stands for the number of relevant prior visits for making clinical decisions for the future. Each hidden state variable captures information about each pair of transitional visits: encounter types, diagnoses, procedures, medications, and duration between two visits. Each observation variable captures information about one or multiple laboratory measurements categorized into appropriate levels such as “normal” and “abnormal.” Figure 1 depicts the modeling framework. In this research, we derive data-driven clinical pathways by computing the most probable sequence of hidden states that corresponds to a sequence of observations in an HMM, using the Viterbi algorithm\(^2\). Maximum likelihood estimate of the parameters in the HMM are obtained using the Expectation Maximization (EM) algorithm\(^4\). In addition, the modeling of the Markov chain state variables facilitates prediction of patients’ next visit. As illustrated in Figure 1, computing the last hidden state variable (in red), which captures the contents of visits \( t+3 \) and \( t+4 \), provides prediction for visit \( t+4 \) (in yellow), the visit that has yet to occur. In other words, it allows us to identify the most likely events to happen in a potentially critical upcoming visit such that patients who may be at risk can be identified early.

Results
Data from 2009 to 2013 on 663 CKD patients were extracted from EHR. Variables include encounter types: office visit and hospitalization; diagnoses: ICD9 codes for CKD stages 1-5, End stage renal disease, Hypertension, AKI, Hyperparathyroidism, Anemia, Proteinuria, Acidosis, Hyperphosphatemia, Glomerulonephritis, Urinary obstruction, Volume depletion, Rhabdomyolysis as diagnoses; drug classes of medications: ACE inhibitors, ARB, Diuretics, Statins, and Proton pump inhibitors (PPI); procedures: Doppler and ultrasound; and laboratory measurements: Albumin, Calcium, Bicarbonate, Creatinine, and Phosphorus. All states and observations in the HMM are categorical variables. Laboratory measures were classified into “low”, “normal”, and “high” according to a current clinical practice guideline in CKD\(^5\). Clinical pathways were generated using patients’ initial 3 visit records. From this dataset, clustering analysis identified 4 subgroups. Within each subgroup, pathway learning and prediction were evaluated using a 5-fold cross validation. Predictions of next visit indicate an accuracy rate of up to 60%, 50% and 72%, depending on if we include as diagnoses all diagnoses mentioned above, CKD alone, or AKI alone, and under a fixed time interval between visits. All other variables from encounter types, drug classes and procures are always included in analysis. When considering episode of AKI as a main outcome, average false positive and negative rates across 4 subgroups are 1% and 30%, indicating our method’s potential but also due to low incidence rate of AKI in certain subgroups. Accurate learning of entire clinical pathway is still underway, and current accuracy is up to 67% for 3-visit pathway when considering AKI only as diagnoses, and all variables from 3 other categories. We anticipate further evaluations and improvements with larger datasets and parameter tuning.

Discussion
Significant diversity and complexity in patients’ EHR data, lack of comprehensive quality standards and performance measures, ad hoc care delivery processes, and the continuously evolving scientific knowledge pose a challenge for efficient pathway learning and prediction. While partly due to the nature of the disease and variability in patients’ conditions, it also urges a review of practices, for treatment variations, inconsistent EHR usage, and improved data capture. Data-driven clinical pathways have the potential to help healthcare providers review their current practices against CPGs, standardize them, and customize them into personalized, patient-centered pathways of care delivery. Future steps include evaluation of the derived pathways by expert clinicians, inclusion of medical costs and critical clinical outcomes into the pathway model, and extension of analytical methods for better accuracy.

References
Accountable Care Organization Provider Network Design:
A systematic data-driven approach

Yuchen Zheng\(^1\), Kun Lin, Ph.D.\(^1\), Kathryn Howard\(^2\),
Thomas White, M.D.\(^3\), Jeremy Pickreign\(^3\), Gigi Yuen-Reed, Ph.D.\(^2\)

\(^1\)IBM T.J. Watson Research Center, Yorktown Heights, NY, USA; \(^2\)IBM Watson Health, Yorktown Heights, NY, USA; \(^3\)Capital District Physicians’ Health Plan, Inc. (CDPHP ®)
Albany, NY, USA

Introduction

In recent years, there has been an increasing emphasis on enabling value-driven health, aimed at improving outcomes, lowering costs, and increasing overall access to care for patients. A prominent example of value-driven delivery systems is the formation of patient-centric Accountable Care Organizations (ACOs). ACOs are groups of physicians, facilities and other healthcare providers, who come together to provide coordinated care to their patients. This study uses prior care patterns to identify groups of provider communities that service the same set of patients. Providers belonging to these naturally-occurring preexisting networks may be more ready to be accountable for managing the health of a population by sharing the risks and benefits of being part of a shared savings program\(^1\), \(^2\).

The proposed method utilizes past care delivery data to detect communities of providers with a higher than expected number of shared patients. We examine the makeup of provider specialties, services, patients, and provider organizations (typically a legal entity composed of multiple providers) related to each community. We prioritize the provider organizations based on their participation and contributions to the patient care in the community in order to find sets of legal entities that are good candidates for shared savings contracts. Communities may share a set of patients for several reasons—professional connections, geographic convenience, or focus on treatment of certain condition(s).

Methods

The data included one year of health insurance claims and enrollment data from a commercial health insurance payer. We study over 100,000 adult patients enrolled in commercial health plans, living in an area where the payer has a large presence. One commercial payer may have limited information about the providers’ interactions and relations to other payers. However, providers often have imperfect health information exchanges, and a payer may be able to provide insight about where their patients receive other services. To focus on patients’ individual providers and specialist referral patterns, we limit our analysis to individual providers (Type I NPIs) and exclude hospitals and pharmacies. We extract the patient identifier, the NPI and the date of the visit from the claims data to construct a mapping between patients and providers. The connections between pairs of providers are weighted by the number of patients the two providers shared. Fast greedy community detection\(^3\) is used on the graph of providers to find communities of NPIs. This is a bottom-up approach, merging sets of NPIs in order to increase the modularity score. This method produces communities of varying size, some of which are too large for contracting purposes. We run the detection algorithm recursively with the requirement that each resulting community has at least 25 NPIs. Any lower bound could be used for community size, but even small numbers of NPIs sharing patients may indicate potential for closer collaboration between their larger organizations.

For contractual reasons, shared-savings programs typically work with provider organizations, the legal entities of the servicing provider groups (taxID). In order to prioritize taxIDs for inclusion in an organization-level ACO, we introduce two key measures for each community and taxID: 1) participation in the community, i.e., the percentage of NPIs contributing to the community; 2) coverage in the community, i.e., the percentage of NPIs contained in the community. For each NPI-level community, we use an F-score based on these two measures. Intuitively, organization-level ACOs approximate the NPI-level communities by maximizing the number of overlapping NPIs while minimizing the number of non-overlapping NPIs. The F-score equation is: 
\[
(1 + 4) \left( \frac{\text{participation\%} \times \text{cover\%}}{4 \times \text{participation\%} + \text{cover\%}} \right)
\]

Results

Considering the 5,500 individual providers servicing over 100,000 patients, our method detected 132 provider communities, ranging from 25 NPIs to over 500 NPIs. Figure 1 shows the dynamics of the recursive community
detection, with each node as one community. The inner circle shows the community candidates from first round of detection, and bigger communities are then further split up into smaller ones. Figure 2 shows an example of how many and which taxID should be contracted together for one community. The F2 Score, which is a weighted measure between participation and coverage, is used to determine the inclusion rules for provider organization.

The resulting communities were heterogeneous across multiple dimensions (geographic location, major conditions of patients, risk scores of patients, % of NPIs with primary care specialties, % of patient care provided by community NPIs). In order to better understand the key factors that drive the formation of such network structure, we explored provider level information for each community, including providers’ affiliation, servicing specialties, and whether providers were primary care providers. To better describe the patient composition in each community, we compiled patient level information (based on the primary care provider of the patient), such as home location, claims cost, conditions, and health risk scores. We segmented patient population by demographic and risk factors to understand the challenges faced by the different communities. Although the literature on what makes a successful ACO is still limited, these traits were examined to determine the strengths and challenges of each community.

Conclusion

We proposed a data-driven method to enable a systematic approach in identifying candidates to form ACO-like provider groups, based on patients’ past utilization history. As much as possible, patients could continue receiving care from their primary care provider and primary care providers could continue making referrals to the same specialists. The payer can now provide evidence of a set of shared patients receiving many of their services from a set of NPIs or taxIDs, in order to increase buy-in to participate in shared savings. We developed key measures to support a payer contracting provider organizations to create shared saving programs. The proposed method will help reduce care disruption in the ACO and have a higher chance to deliver high quality and coordinated care. These programs can help providers succeed by preparing them for a market moving toward value-based care.

References

Automatic identification and extraction of design patterns of EHR-driven phenotyping algorithms

Yizhen Zhong, BS1; Luke Rasmussen, MS1; Justin Starren, MD, PhD1; Yuan Luo, PhD1
1Northwestern University, Feinberg School of Medicine, Chicago, IL

Introduction: The electronic Medical Records and Genomics (eMERGE) network [1] is an NHGRI-sponsored initiative to further the development and implementation of EHR-derived phenotypes across multiple institutions. The phenotyping algorithms developed, validated and implemented through multi-site collaboration are primarily recorded as text documents [2] and are re-implemented in a computable format at each site. There is a need to apply best practices to ensure consistent implementations derived from narratives. This is complicated not only by the nuances of how data is collected within the EHR, but also by the heterogeneity across EHR vendors, implementations and individual use. Awareness of these differences and accounting for them up front in the phenotype development process may increase the efficiency of the development process, and the accuracy of the resulting phenotype.

Methods: To facilitate consistent and reproducible phenotyping algorithm development, we previously proposed “phenotype design patterns” [3]. A phenotype design pattern is intended to address a commonly seen issue in the use of EHR data, and provide guidance to a solution. The initial proposed set of phenotype design patterns was developed by a review of existing phenotypes, and leveraged expert opinion. The process was time-consuming, and would not scale as more phenotype algorithms are defined. Thus, automated approaches to identify and extract candidate patterns would advance our understanding on best practices in the field of phenotyping. Using the narrative description of each phenotype algorithm downloaded from Phenotype KnowledgeBase (PheKB.org) with a status of Final or Validated [3], one of the authors (LVR) extracted fragments that described aspects of the data elements and/or logic used in the algorithm. In this pilot study, we focused on the phenotype design patterns as shown in Table 1. The data consists of 160 fragments of 2290 words in total extracted from 41 phenotyping algorithms. We substituted all the numbers with _number, removed stop words, and stemmed all remaining words. We experimented with features including bag-of-words, sites of the phenotyping algorithms, targeted conditions (e.g., T2DM), UMLS concept unique identifiers (CUIs), and UMLS semantic types (STs). The CUIs and STs were obtained by running MetaMap [4] over the text fragments, and are included to battle the word sparsity problem. The site and condition features are included to account for possible site- and condition associated biases. Observing that synonyms may occur frequently in narratives, we also experimented with word embedding that was trained on a large corpus of EHRs from the MIMIC II dataset using the word2vec toolkit [5] with the embedding dimension set to 200. We performed classification and clustering of design patterns using features above-mentioned to demonstrate the feasibility of automatic identification and extraction of design patterns. As one sentence may have multiple labels, we formulated the classification problem as building multiple one-versus-rest classifiers. We split the data into a training set and a test set according to a 7:3 ratio stratified by the percentage of fragments for each pattern. We trained a linear kernel support vector machine (SVM) with parameters C and l1/l2 regularization tuned by cross validation on the training set. The number of fragments for clustering was 126 with the removal of fragments with multiple labels. The clustering was performed using k-means algorithm (k=8).

Results: The classification and clustering performance was evaluated on the held-out test set using f1score for each class and macro/micro averaged f1 for overall assessments as shown in Table 2 and Table 3. The tables show combinations of different types of features with either bag-of-words model or embedding model. The maximum micro averaged f1 is 0.73 for classification using features of bag-of-words and condition. We obtained 0.49 micro averaged f1 for clustering using bag-of-words model. The overall performance demonstrates moderate evaluation scores suggesting the practicability of automatic identification and extraction of phenotyping design patterns.

Discussion: From Table 2, it does seem that the site and condition preferences may add to the classification accuracies of design patterns including “Disease”, “Dates”, “Rule of N” and “Evidence” as reflected by the increased f1. It is also interesting that embedding, by capturing the word semantic relatedness, leads to increased f1 for the “Med”. For example, the fragments “there was no indicator of topical or cream in the flanking text” which was incorrectly classified using bag-of-words model was better represented by embedding model as key words “topical” or “cream” are semantically related to “medic”. From Table 3, embedding model outperformed bag-of-words model for “Dates” and “Evidence” classes which may because for example the word pair “separate” and “distinct” from one fragment has similar embedded vector representations to word pair “days” and “dates” from another fragment. Without class specific feature engineering, the overall moderate performance shows the promise of automated classification and clustering
ing of phenotyping design patterns. We are extending our algorithm to extracting more design patterns from phenoty pe algorithms.

Table 1. Phenotyping design patterns and descriptions.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Description</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dates</td>
<td>When requiring a count of items, make sure they happen on multiple dates, possibly with some time interval between them.</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>Disease</td>
<td>Make sure the patient has been in to see a doctor to be screened for a condition. This may also apply to labs, to ensure that a lab value was checked &amp; came back normal.</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>Where</td>
<td>Knowing if something was inpatient or outpatient is important. Kind of like Transient Conditions, if the patient is in the hospital, we exclude data in many cases.</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Evidence</td>
<td>Codes have a severity associated with them; with NLP, consider adjectives that describe the level of evidence and certainty; with diagnosis codes, require that the code be the primary diagnosis.</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Rule of N</td>
<td>More evidence is often required, especially when recurring codes gives a higher level of certainty that a condition exists and wasn't just a rule-out.</td>
<td>39</td>
<td>14</td>
</tr>
<tr>
<td>Credentials</td>
<td>If you need a physician to make a diagnosis, make sure a physician entered it. Likewise, if you need a specialist to make the diagnosis, ensure that is the data you are pulling.</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Negation</td>
<td>Determine if negated mentions of terms exist. In some instances, you may need to confirm a negative mention exists. In others, you may need to filter out terms that are negated.</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>Med</td>
<td>When checking for medications, sometimes you must look at parts of the medication, such as specific brand name, groups of medications, ingredient, how it was delivered (topical vs. oral)</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2. F1 scores for phenotyping design pattern classification. Abbreviations used: BOW: bag-of-words; CUI: UMLS concept unique identifiers; ST: UMLS semantic types; Cond: targeted condition of phenotyping algorithm. Bold indicate best results.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Med</th>
<th>Disease</th>
<th>Rule of N</th>
<th>Dates</th>
<th>Evidence</th>
<th>Credential</th>
<th>Where</th>
<th>Negation</th>
<th>Macro-fl</th>
<th>Micro-fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>0.29</td>
<td>0.71</td>
<td>0.65</td>
<td>0.67</td>
<td>0.55</td>
<td>1</td>
<td>0.73</td>
<td>0.69</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>BOW +Site</td>
<td>0.33</td>
<td>0.83</td>
<td>0.75</td>
<td>0.86</td>
<td>0.4</td>
<td>0.4</td>
<td>1</td>
<td>0.73</td>
<td>0.71</td>
<td>0.7</td>
</tr>
<tr>
<td>BOW +Cond</td>
<td>0.6</td>
<td>0.4</td>
<td>0.79</td>
<td>0.8</td>
<td>0.89</td>
<td>0.55</td>
<td>0.8</td>
<td>0.67</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>BOW +CUI</td>
<td>0.57</td>
<td>0.4</td>
<td>0.58</td>
<td>0.82</td>
<td>0.67</td>
<td>0.75</td>
<td>0.91</td>
<td>0.55</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>BOW +ST</td>
<td>0.4</td>
<td>0.21</td>
<td>0.65</td>
<td>0.78</td>
<td>0.4</td>
<td>0.67</td>
<td>1</td>
<td>0.67</td>
<td>0.65</td>
<td>0.61</td>
</tr>
<tr>
<td>Embedding</td>
<td>0.67</td>
<td>0.56</td>
<td>0.61</td>
<td>0.71</td>
<td>0.67</td>
<td>0.67</td>
<td>0.62</td>
<td>0.46</td>
<td>0.66</td>
<td>0.62</td>
</tr>
<tr>
<td>Embedding+Site</td>
<td>0.67</td>
<td>0.4</td>
<td>0.72</td>
<td>0.69</td>
<td>0.4</td>
<td>0.67</td>
<td>0.67</td>
<td>0.36</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Embedding+Cond</td>
<td>0.75</td>
<td>0.43</td>
<td>0.63</td>
<td>0.55</td>
<td>0.8</td>
<td>0.67</td>
<td>0.57</td>
<td>0.64</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Embedding+CUI</td>
<td>0.67</td>
<td>0.55</td>
<td>0.67</td>
<td>0.74</td>
<td>0.86</td>
<td>0.67</td>
<td>0.89</td>
<td>0.5</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>Embedding+ST</td>
<td>0.67</td>
<td>0.46</td>
<td>0.67</td>
<td>0.52</td>
<td>0.4</td>
<td>0.67</td>
<td>1</td>
<td>0.5</td>
<td>0.67</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 3. F1 scores for phenotyping design pattern clustering.

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Med</th>
<th>Disease</th>
<th>Rule of N</th>
<th>Dates</th>
<th>Evidence</th>
<th>Credential</th>
<th>Where</th>
<th>Negation</th>
<th>Macro-fl</th>
<th>Micro-fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>0.74</td>
<td>0.44</td>
<td>0.22</td>
<td>0.64</td>
<td>0.3</td>
<td>0.76</td>
<td>0.53</td>
<td>0.47</td>
<td>0.52</td>
<td>0.49</td>
</tr>
<tr>
<td>BOW +Site</td>
<td>0.22</td>
<td>0.09</td>
<td>0.38</td>
<td>0.42</td>
<td>0.28</td>
<td>0.41</td>
<td>0.21</td>
<td>0.21</td>
<td>0.32</td>
<td>0.29</td>
</tr>
<tr>
<td>BOW +Cond</td>
<td>0.52</td>
<td>0.25</td>
<td>0.19</td>
<td>0.61</td>
<td>0.32</td>
<td>0.52</td>
<td>0.46</td>
<td>0.53</td>
<td>0.45</td>
<td>0.4</td>
</tr>
<tr>
<td>BOW +CUI</td>
<td>0.26</td>
<td>0.33</td>
<td>0.54</td>
<td>0.67</td>
<td>0.41</td>
<td>0.67</td>
<td>0.61</td>
<td>0.52</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>BOW +ST</td>
<td>0</td>
<td>0.39</td>
<td>0.3</td>
<td>0.67</td>
<td>0.36</td>
<td>0.86</td>
<td>0.46</td>
<td>0.33</td>
<td>0.45</td>
<td>0.41</td>
</tr>
<tr>
<td>Embedding</td>
<td>0.38</td>
<td>0.15</td>
<td>0.41</td>
<td>0.7</td>
<td>0.53</td>
<td>0.47</td>
<td>0.21</td>
<td>0.47</td>
<td>0.45</td>
<td>0.38</td>
</tr>
<tr>
<td>Embedding+Site</td>
<td>0.38</td>
<td>0.16</td>
<td>0.38</td>
<td>0.42</td>
<td>0.13</td>
<td>0.41</td>
<td>0.2</td>
<td>0.32</td>
<td>0.32</td>
<td>0.3</td>
</tr>
<tr>
<td>Embedding+Cond</td>
<td>0.2</td>
<td>0.15</td>
<td>0.48</td>
<td>0.61</td>
<td>0.48</td>
<td>0.36</td>
<td>0.46</td>
<td>0.44</td>
<td>0.42</td>
<td>0.41</td>
</tr>
<tr>
<td>Embedding+CUI</td>
<td>0.34</td>
<td>0.08</td>
<td>0.51</td>
<td>0.67</td>
<td>0.45</td>
<td>0.3</td>
<td>0.56</td>
<td>0.57</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>Embedding+ST</td>
<td>0.33</td>
<td>0.3</td>
<td>0.48</td>
<td>0.64</td>
<td>0.24</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.44</td>
<td>0.41</td>
</tr>
</tbody>
</table>

References

An Error Analysis of Dictated Clinical Documents at Different Processing Stages

Li Zhou, MD, PhD1,2,3, Warren W. Acker1, Adam B. Landman, MD1,2,3, Evgeni Kontrient, MD4, Raymond Doan1, Suzanne V. Blackley1, David Mack5, David W. Bates, MD, MSc1,2,3, Foster R. Goss DO, MMSc5

1Brigham and Women’s Hospital, Boston, MA; 2Partners HealthCare, Boston, MA; 3Harvard Medical School, Boston, MA; 4North Shore Medical Center, Salem, MA; 5University of Colorado, Aurora, CO

Introduction

Clinical documentation represents one of the most time-consuming and costly parts of using an electronic health record (EHR) system. Speech recognition (SR), the computer recognition of spoken language into text, has been a promising technology for clinical documentation since the 1980s. Recent studies reported that 90% of hospitals plan to expand their use of SR technology. SR can assist clinical documentation at the back end, where a clinician uses a telephone to dictate, the captured recording is then recognized by an SR engine, and then the SR generated text is edited by a medical transcriptionist before sending it back to EHR for review and signature. However, issues including documentation quality, productivity, and physician satisfaction have not been widely studied. This abstract presents findings from part of a larger ongoing study and focuses on analyzing errors generated by back-end SR.

Methods

Our preliminary dataset includes a stratified random sample of 50 dictated notes from Brigham and Women’s Hospital and 10 notes from the University of Colorado Hospital. Both institutions use the same transcription vendor. The note types include clinic notes, operative notes, progress notes and discharge summaries. We reviewed each note at each of the four processing stages of dictation, namely 1) the original audio file that was dictated by the provider, 2) the note generated by SR engine of the vendor transcription service (SR note), 3) the note edited by a professional medical transcriptionist (MT note), and 4) the final note reviewed and signed by a clinician (SN note). In order to create a gold standard note, under a practicing physician’s supervision, a PharmD candidate or a medical student listened to the original dictation, edited the MT note to reflect that dictation, and also confirmed medical information with the patient’s chart.

To analyze errors we used Knowtator, an open-source annotation tool to annotate differences from the gold standard in notes at different processing stages. We adopted the annotation scheme of Zafar et al and modified it after consultation with a speech recognition expert and our experience with a small sample of notes. Our annotation scheme was broken into three levels: 12 general error types (e.g., insertion and enunciation), 14 semantic error types (e.g., medication and diagnoses), and clinically significant errors. Errors were considered clinically significant if they could potentially impact clinical care and these errors were validated by physician researchers.

We determined the length of time to dictate a note and the turnaround time for each note version. We analyzed differences in the SR note, MT note, and SN note from the gold standard. We determined error rate (number of errors divided by the number of words), the percentage of each error type by overall errors (number of errors of that error type divided by the total number of errors), and the percentage of notes with at least one clinically significant error (number of notes with one or more clinically significant error divided by the total number of clinical notes). We repeated these analyses for SR, MT and SN notes, for each note type, and across all note types.

Results

From BWH, 41 clinic notes, 5 operative notes, and 4 progress notes were analyzed; UCH provided 5 operative notes and 5 discharge summaries. Each note on average contained 516 words (median: 395 words; range: 44-1790 words). Average dictation time was 5 minutes (median: 4 minutes; range: 21 seconds-31 minutes). On average, the turnaround time for a dictated note was 4 hours (median: 1.3 hours; range: 10 minutes-23 hours) from completion of the dictation to upload into the EHR system.

Detailed results of our error analysis were shown in Table 1. In SR notes, there were 1631 errors in total and 27 errors per note, with an overall error rate at 6.9% and 87% of notes containing errors. There were 515 (31.6%) deletion errors, 444 (27.2%) enunciation errors and 335 (20.5%) insertion errors. We identified 203 (12.5%) errors were about clinical information, including 52 (3.2%) diagnoses, 41 (2.5%) symptoms, and 38 (2.3%) medications. 66 (4.1%) errors were judged to be clinically significant and 24 (40%) notes had at least one clinically significant error.
In MT notes, there were 88 errors in total and 1.5 errors per note, with an overall error rate at 0.4% and 47% of notes containing errors. There were 30 (34.1%) enunciation errors and 29 (33.0%) deletion errors. We found that 13 (14.8%) errors were about clinical information, including 4 (4.6%) diagnoses, 4 (4.6%) procedures, 2 (2.3%) symptoms, and 1 (1.1%) medication. Four (4.6%) errors were judged to be clinically significant and 4 (6.7%) notes had at least one clinically significant error.

In SN notes, there were 27 errors in total and about 1 error every two notes, with an overall error rate at 0.1% and 20% of notes containing errors. There were 10 (37.0%) insertion errors and 8 (19.6%) enunciation errors. Five (18.5%) errors were of clinical information, including 2 diagnoses, 1 lab, 1 image and 1 symptom. Three (11.1%) errors were judged to be clinically significant and 3 (5%) notes had at least one clinically significant error.

Table 1. Error Rates and Error Types in Dictated Documents by Note Type and Processing Stage

<table>
<thead>
<tr>
<th>Note Type (n)</th>
<th>Note Stage</th>
<th>Total Errors n (%)</th>
<th>Generic Errors n (%)**</th>
<th>Semantic Errors n (%)**</th>
<th>Clinically Significant Errors n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Deletion</td>
<td>Enunciation</td>
<td>Insertion</td>
</tr>
<tr>
<td>Clinic Notes (41)</td>
<td>SR</td>
<td>1550 (9.8)</td>
<td>371 (23.2)</td>
<td>330 (20.7)</td>
<td>219 (14.0)</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>68 (0.6)</td>
<td>24 (35.3)</td>
<td>25 (36.8)</td>
<td>10 (14.7)</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>17 (0.1)</td>
<td>4 (23.5)</td>
<td>7 (41.2)</td>
<td>4 (23.5)</td>
</tr>
<tr>
<td>Operative Notes (10)</td>
<td>SR</td>
<td>232 (3.3)</td>
<td>56 (24.1)</td>
<td>52 (22.4)</td>
<td>71 (30.6)</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>16 (0.2)</td>
<td>5 (31.3)</td>
<td>2 (12.5)</td>
<td>6 (37.5)</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>9 (0.1)</td>
<td>1 (11.1)</td>
<td>1 (11.1)</td>
<td>6 (66.7)</td>
</tr>
<tr>
<td>Progress Notes (4)</td>
<td>SR</td>
<td>133 (5.9)</td>
<td>46 (34.6)</td>
<td>35 (26.3)</td>
<td>23 (17.3)</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>4 (0.2)</td>
<td>0 (0.0)</td>
<td>3 (75.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>11 (0.1)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Discharge Summaries (5)</td>
<td>SR</td>
<td>116 (5.2)</td>
<td>42 (36.2)</td>
<td>27 (23.3)</td>
<td>22 (19.0)</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>11 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>All Notes (60)</td>
<td>SR</td>
<td>1631 (6.9)</td>
<td>515 (31.6)</td>
<td>444 (27.2)</td>
<td>335 (20.5)</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>88 (0.4)</td>
<td>29 (33.0)</td>
<td>30 (34.1)</td>
<td>16 (18.2)</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>27 (0.1)</td>
<td>5 (18.5)</td>
<td>8 (29.6)</td>
<td>10 (37.0)</td>
</tr>
</tbody>
</table>

*% is number of errors and % is error rate which is defined by total number of errors divided by the total number of words in the notes. ***% is the number of errors of that specific type divided by the total number of errors. **% is defined as the number of notes containing at least one clinically significant errors divided by the number of notes.

Discussion

Our results showed that back-end dictation service had a relatively quick turnaround time and a low error rate. While many errors were generated by SR, most (~94%) were corrected by the medical transcriptionist manually. This suggests that the addition of a human editing a SR-generated note is invaluable. A recent study of EHR related malpractice cases found that 20% were due to incorrect information in the EHR, although the percentage due to dictation is unknown.

Errors still left in the signed note suggests that some providers may not review their dictated notes thoroughly or at all. For example, we also found that 3% of signed notes contained a blank space the transcriptionist marked as “??___??” (e.g., one was on the dosage of Vicodin), indicating they were unsure of the word and for the physician to fill it in. These cases indicate if physicians use SR directly, they may have to spend a considerable amount of time correcting the SR-generated text, although the errors should be less than our results as the SR can be trained by the individual physician. Automated error detection may help improve the accuracy of dictated documents.

Conclusion

Clinical dictation systems using back-end SR technology have a quick turnaround. SR-generated documents contain a high error rate and clinically significant errors. Fewer errors in signed documents show that medical transcriptionists and human editing play a critical role in quality assurance.

Acknowledgements: This study was funded by AHRQ grant R01HS024264. We thank Kenneth Lai, Maxim Topaz, Victor Lei, Yoobin Kim, and Marie Meteer for their help with this study.

References
Protecting patient privacy in cyber environments

Panelists
Samantha Adams, PhD¹, Bonnie Kaplan, PhD, FACMI², Paul R DeMuro, JD, PhD³,⁴, Tony Solomonides, MSc⁵

Moderator
Jos Aarts, PhD, FACMI⁶

¹University of Tilburg, The Netherlands; ²Yale University, New Haven, CT; ³Broad and Cassel, Attorneys at Law, Fort Lauderdale, FL; ⁴Nova Southeastern University, Fort Lauderdale, FL; ⁵NorthShore University HealthSystem, Evanston, IL; ⁶State University of New York at Buffalo, NY

Abstract
Confidentiality in the medical encounter is crucial to providing adequate patient care. Health data is therefore privileged and protected by legal mechanisms. Health systems use electronic records and large-scale databases. Increasingly consumers use also IT to collect, store and share data about daily life and health behaviors. Sharing data via network-based systems or storing it ‘in the cloud’ produces multiple ‘digital selves’, health ‘data doubles’ and ‘virtual patients.’ With so many stakeholders involved much data is produced without clear governance structures, blurring the view of what is done with the data. These problems are exacerbated through the networked, distributed nature of health data collection and convergence of protected hospital systems, commercial collection and aggregation of data and consumer health technologies. This brings patient privacy into the realm of cybersecurity. This panel explores how cybersecurity impacts the governance of critical IT infrastructures and mitigation of threats, what sociotechnical challenges are related to protection of large-scale HIT systems, how surveillance and bioethics studies seek to understand threats to personal privacy in the context of networked technologies and finally what changes to laws and regulations would be required.

Keywords: Patient privacy; data protection; cybersecurity; legal, ethical and regulatory issues; governance

Introduction
“If you need to be convinced that you’re living in a science-fiction world, look at your cell phone,” writes Bruce Schneier in the introduction of his book Data and Goliath [1]. He paints a picture of how massive amounts of data are gathered and the public’s perception of privacy is changing drastically. The public accepts that in exchange for being connected, service providers know at all times where they are and what they do. Health data are not excluded. In return for services, consumer exchange and store their data on network-based systems or ‘in the cloud’ and create digital identities, for example as ‘virtual patients.’ Often governments and private businesses make a distinction between data and metadata. Data is about content, and metadata is about context. However, as NSA and CIA director Michael Hayden said, “Metadata tells you everything about somebody’s life.” Schneier argues that the difference is largely illusory [1].

Data collection and use can threaten individual privacy. Much data is produced without clear governance structures and stipulations concerning who will use which data and to what end. Lack of such structures and lack of transparency about those that do exist, result in patients, professionals and institutions having an insufficient view of what is done with the data [2]. Another risk includes healthcare information systems’ infamous weaknesses that result from complex code and software insecurity [3]. These problems are exacerbated through the networked, distributed nature of health data gathering and use, especially increasing convergence between protected hospital systems, commercial collection and aggregation of data, and consumer health technologies, such as health apps on mobile devices and wearable sensors.

Most institutions are insufficiently prepared to handle system shutdown resulting from internal network malfunction or external attack on critical infrastructures, bringing issues of data protection and patient safety into the sphere of cybersecurity. Directed attacks on both commercial and institutional systems threaten the privacy of
individuals’ personal data [4]. Medical identity theft is already an economic and social concern, and disclosures about personal health can have other consequences for individuals [2, 5].

At the same time, public perceptions of privacy and security are changing. This is evident in the push toward more ‘publicness’ of health data and new pay-for-value and cost-effectiveness reimbursement structures that require a greater sharing of data and health information exchange [6]. Moreover, privacy is not absolute, but rather a relational property; individuals may be more or less willing to trade off privacy depending on the situation and perceived benefits [1,7]. Shifts in the boundary between public and private, together with more use of distributed, networked technologies in clinical practice and research have led to concrete legal, ethical and infrastructural barriers and challenges related to computational agents, transmission protocols, care practices and legal tools.

Currently, five separate arenas of study address governance issues related to cybersecurity, privacy, and the protection of (health) data: (1) cybersecurity studies examine the socio-political and technical governance of critical infrastructures and mitigation of threats to systems such as botnets; (2) patient safety examines concrete sociotechnical and organizational challenges related to the protection of large-scale HIT systems; (3) surveillance studies and (4) studies in bioethics each examine different aspects of (potential) threats to personal privacy through networked technologies, and (5) legal studies examine (proposed) changes to the laws and regulations that govern data creation and exchange.

The panel

This panel explores how governance structures for protecting patient privacy in cyber environments are translated into concrete laws and regulations, institutional policies and procedures, and technical protocols for health practice. More specifically the panelists will explore following questions:

- What are current ethical-legal challenges and how are they related to practice and governance?
- What legal and regulatory changes are required to insure patient privacy in a cyber context?
- How can institutions safeguard integrity, privacy and confidentiality in a cyber context?

The panelists

**Samantha Adams**, PhD is associate professor of eHealth Governance and Regulation at the Tilburg Institute for Law, Technology and Society (TILT) in Tilburg, the Netherlands. Adams has more than 15 years’ experience in researching ethical, politico-legal and social issues related to the use of HIT. She has advised both government and healthcare institutions in the Netherlands about responsible implementation and use of networked systems for health and currently advises three European-based research projects. She is a multiple-term appointee to the Ethics Committee of the American Medical Informatics Association (AMIA) and an active member of three working groups. She is an expert in innovative qualitative research methods. In 2015, Dr. Adams led a comparative review project of Cybersecurity Governance in 5 countries commissioned by the Dutch Ministry of Security and Justice. Dr. Adams will briefly discuss the current state of research on the cybersecurity of critical infrastructures and the place of healthcare in this research. She will then outline pending changes to the General Data Protection Regulation in Europe and the consequences of these changes for healthcare institutions. She will conclude by showing how these changes are operationalized in practice by health institutions and government regulatory bodies in the Netherlands and reflecting on what lessons can be learned from these practices.

**Bonnie Kaplan**, PhD, FACMI, of the Yale Center for Medical Informatics, is a Yale Interdisciplinary Bioethics Center Scholar, and a Faculty Fellow of the Yale Law School’s Information Society Project. She has long experience in evaluating effects of health information technology on people and organizations. Her research addresses informatics ethical, legal, and privacy issues; user perspectives and experiences with health information technology; and ethnographic sociotechnical evaluation. Most relevant to this panel, she recently published papers on health data privacy in *The Cambridge Quarterly of Healthcare Ethics*, and the paper she co-authored on ethical issues in telemedicine has been among the most cited since its publication there in 2008. She is a past chair of the American Medical Informatics Association’s (AMIA) Ethical, Legal, and Social Issues Working Group; AMIA’s People Organizational Issues Working Group; and the International Medical Informatics Association’s (IMIA) Organizational and Social Issues Working Group. She is an elected Fellow of the American College of Medical Informatics. Dr. Kaplan will address ethical-legal challenges concerning privacy and security for clinical and patient-produced data, especially identifying problems in US regulation and gaps in public understanding of privacy protection. She will suggest changes to better protect patient privacy.

**Paul R. DeMuro**, JD, PhD is a health information technology attorney with over 35 years of health law experience. He is Of Counsel to the law firm of Broad and Cassel, based in Fort Lauderdale, and Associate Professor in the College of Pharmacy of Nova Southeastern University in Fort Lauderdale. He is Chair of the
Finance and Investment Committee of the American Medical Informatics Association and a member of its Ethical, Social and Legal Issues Working Group. His research areas include the use of health information technologies to improve quality and reduce costs in value-based purchasing, and the legal considerations involved. Dr. DeMuro will discuss a number of the legal considerations involved in health systems with respect to cybersecurity, including HIPAA privacy and security, patient privacy, data protection and governance, particularly those arising from hacking, and how to protect from hacking. He also will address institutional preparedness, including education of employees, implementation of a data backup plan, retention schedules, incident response procedures, access rights, and anti-malware tools. Finally, he will address the importance of compliance programs in this context.

Tony Solomonides, MSc is Director of Clinical Research Informatics at NorthShore University HealthSystem, Evanston. After a twin-track academic career as a biomedical informatician and as a computer scientist, culminating in his work in ‘health grids’ and serving as Vice-President and Chair of the Scientific Board of HealthGrid from 2006 to 2010, Tony was invited to join NorthShore, in late 2011. In this position, he has contributed to the design and analysis of clinical data in a number of projects and is now co-PI on the PCORI-funded CAPriCORN clinical data research network in the Chicago area. He is chair-elect of AMIA’s Ethics, Legal and Social Issues Working Group. Tony’s broader vision is to enhance the role of the patient as ‘expert’ in his or her own condition and in maintaining contact both as a means of improving accuracy and of continuing informed consent for the use of healthcare data for research. Safeguards of the integrity, privacy and confidentiality of the data would be maintained through formal ‘norms’ and annotation-enriched data. Tony will discuss the potential of data annotation, the possibility of norm- rather than rule-based privacy management, and use of the concepts of formal institutions and agents.

Jos Aarts, PhD, FACMI is associate professor of biomedical informatics at the University at Buffalo and assistant professor in the Institute of Health Policy and Management of Erasmus University Rotterdam. His research interests are the sociotechnical perspective of health IT in relation to professional work practices and workflow, usability and safety of HIT. He is an elected fellow of the American College of Medical Informatics. He is a past-chair of AMIA’s People and Organizational Issues Working Group. Following the panelist’s presentations, the moderator will start the discussion by asking a few questions to help the audience understand the similarities and differences among the various perspectives. He will then open the floor for a free-ranging discussion of all aspects of protecting patient privacy in cyber environments.

Learning objectives

After participating in the session, the attendant should be able to:

• Describe institutional practices that safeguard against external and internal threats to network-based health systems and breakdowns in networked information systems;
• Identify current ethical and legal challenges related to protecting patient privacy in cyber environments;
• Reflect on areas of improvement for governance structures and policy for protecting against cyber attacks.

Conflict of interest

The participants have no conflicts of interest to declare.

References

The Human Microbiome: Informatics Challenges And Opportunities

Alexander V. Alekseyenko, PhD¹, Michael J. Becich, MD, PhD, FACMI², Todd Z. DeSantis, MS³, Jack Gilbert, PhD⁴, Georg K. Gerber, MD, PhD, MPH⁵
¹Program for Human Microbiome Research, The Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC; ²Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA; ³Informatics Division, Second Genome, Inc., South San Francisco, CA; ⁴Department of Surgery, University of Chicago, Chicago, IL; ⁵Massachusetts Host-Microbiome Center and Center for Advanced Molecular Diagnostics, Department of Pathology, Brigham and Women’s Hospital at Harvard Medical School, Cambridge, MA

Abstract

Human microbiome, the collection of all microorganisms cohabiting the human body, is making an impact on a broad range of biomedical fields. Actionable associations of the host microbiota with health conditions span many health areas, such as oral and dental health, inflammatory bowel disease, metabolic syndrome, and other systemic conditions, and cancers. The technologies for microbiome science are inherently heavy on informatics. Many leading microbiome scientists would admit inadequate involvement of informatics experts is currently the key obstacle to progress. The informatics-related challenges and opportunities include issues with representation of multi-species data, creation and validation of reference databases for genes and genomes, benchmarking and validation of computational and analytic techniques for microbiome data upstream processing and downstream analysis, and integration of microbiome data with clinical data and other multi-omic data both for research and at bed side. In addition, the White House Precision Medicine Initiative specifically calls the biobanking and sharing of microbiome cohorts an important characteristic that will enable health care providers to tailor treatment and prevention strategies to individual patients. The new National Microbiome Initiative announced in May 2016, is further underlying the national importance of this area. This panel will stimulate the development of specific action plan to address the challenges and capitalize on opportunities for informaticists in translational microbiome science.

Panel description

Microbiome is a term used to describe the collection of all microorganisms that colonize our bodies. Early estimates posed 1-to-10 overabundance of microbial cells to human cells in a typical human body. These have recently been corrected towards parity, which is still overwhelming in terms of the potential for functional interactions between the host and the microbe. Whereas understanding the human genome has received enormous amount of attention in the last 15 years, the impact of host associated microbiome genomes (the metagenome) and its function in human health and disease is still in very early stages of being conceptualized. The recent surge in microbiome science has been afforded by increasing availability of technologies, such as high-throughput DNA sequencing, mass spectrometry, automated cell sorting, etc. Most notably, of course, the ability to sequence hundreds of individual microbiomes in a single run of a sequencing instrument has resulted in a perfect storm of microbial community profiling datasets addressing a vast array of biomedical questions. Most notably microbiome has made contributions in the areas of inflammatory bowel disease, C. difficile infection (CDI) management, the metabolic syndrome and pre-diabetes. It is the better understanding of the interplay between the host and its microbiota that has allowed for innovative fecal transplant treatments to be successfully used to treat CDI. Increasingly, microbiome is being utilized for hospital associated infection management and more recently to build personalized precision medicine approaches to clinical decision support.

The complexity of the microbiome itself and the sophistication of the microbiome assay technologies necessitates involvement of quantitative scientists in many branches of knowledge including informatics. Many data challenges are still outstanding in the microbiome arena. For example, there is currently a gap in standards and standard reference resources for microbial identification. This results in lack of comparability of outcomes of different experimental groups. Poor reproducibility of results is a major barrier to translating the findings into useful application. Another translational roadblock is the fact that the experimental microbiome data is typically far removed from the health record data. A standard for integration of microbiome assay results into clinical reports does not currently exist.

This panel brings together experts in microbiome science and biomedical informaticists to discuss a plan to capitalize on the opportunities to improve healthcare using the vast biological knowledge afforded by studies of the microbiome.
Brief description of each panelist’s presentation

- Dr. Alekseyenko, “The Human Microbiome in Health and Disease”
  Technology has enabled characterization of the host and microbiome physiological state on many dimensions in essentially panomic way. These technological advances in microbiome research have enabled discoveries across the biomedical spectrum. In this presentation, we overview the promise that microbiome carries for novel discoveries in key biomedical areas.

- Dr. Becich, “The Role of Computational Pathology and Biomedical Informatics in Microbiomics and Precision Medicine”
  Computational Pathology is the academic discipline that has grown from Pathology Informatics community. The Association for Pathology Chairs has declared the development of computational pathology as critical to the academic research mission of forward thinking Medical School Departments. This presentation will discuss the synergies between biobanking, computational pathology and biomedical informatics in the microbiome era and where we are headed in building programs at the University of Pittsburgh to support a Center of Excellence in Microbiomics, which will greatly serve the mission of the Precision Medicine Initiative Cohort.

- Mr. DeSantis, “Selecting Therapeutic Candidates from Mucosal Biospecimens Using Network Analyses”
  Microbiome insight comes from integrating host transcriptomics, microbe meta-transcriptomics, metabolomics and 16S analyses, often without a definitive Rosetta Stone. Integration of these new data with existing public and proprietary curated host-side interaction networks presents many challenges in discovering actionable biomarkers. This presentation describes how to leverage these host-side and similar microbe-side resources and how to set up validation screens.

- Dr. Gilbert, “The Hospital Microbiome Project: Ecology and Epidemiology”
  The bacteria, fungi and viruses that are shed from our bodies create a dynamic ecosystem in our built environment. The hospital environment is no different, and was characterized in our study by examining how microbes found associated with patients and staff, colonized and persisted in the hospital rooms and nursing stations. This dataset suggests that the microbial community found in the patient’s room, and the microbial interactions between patients and staff likely influence patient health outcomes, and we therefore propose that the microbial environment of the patient should be taken into consideration when considering treatment strategies.

- Dr. Gerber, “Prediction of Patient Outcomes from Longitudinal Microbiome Data”
  The microbiome is inherently dynamic, changing in response to the host, the environment, and interactions among the microbes. This presentation will highlight the challenges of designing and analyzing longitudinal microbiome studies. Longitudinal studies currently being conducted at the Massachusetts Host-Microbiome Center investigating the role of the microbiota in food allergy and in Clostridium difficile recurrence will be discussed as examples. Progress on new computational methods that specifically address the challenges of analyzing longitudinal microbiome data, including sparse and irregular temporal sampling, noisy count-based data, phylogenetic structure, complex dynamics, unmeasured host and environmental processes, and model interpretability, will also be discussed.

Relevance of topic and anticipated audience:
The White House precision medicine initiative, which is likely to shape the research funding policy in the next few years, specifically calls microbiome an important characteristic for Precision Medicine. Recently the White House Office of Science and Technology Policy has announced the launch of the National Microbiome Initiative, which aims to foster the integrated study of microbiomes across different ecosystems. The original call for information prior to the launch has specifically indicated that informatics involvement is key to microbiome science by declaring that “The development of platform technologies, reference libraries, and databases useful for microbiome research in all habitats” is an example of work relevant to OSTP goals. The final announcement calls for new tools to and technologies to develop knowledge necessary to “manage microbiomes in a manner that prevents dysfunction or restores healthy function”. We anticipate that this panel will be interesting to attendees working in the translational informatics realm, specifically those focusing on precision medicine.
Discussion questions:
1. What resources to support translational microbiome science currently exist?
2. What key expertise is lacking or inadequately present on the microbiome arena?
3. What translational value does the microbiome hold and in what specific areas of medicine and health?
4. What are the national and regional centers for microbiome research and what opportunities for involvement exist?
5. How is microbiome science incorporated in biomedical informatics training?

Learning objectives:
1. Recognize the technologies exist to survey the human microbiome and the biological aspects they capture;
2. Identify areas where microbiome has made an impact on patient treatment, hospital policies, or understanding disease processes;
3. Identify areas with existing opportunities for improving care, policies, and knowledge using microbiome science;
4. Identify informatics challenges for wide and uniform adoption of microbiome science in healthcare.

The organizer confirms that all participants have agreed to take part on the panel.
Who Watches the Watchers: tools and practices for monitoring and measuring CDS performance and assessing CDS effectiveness and value
Organizer: Jan Marie Andersen, MA

Moderator:
Adam Wright, PhD\textsuperscript{a,b,c}

Panelists:
Jan Marie Andersen, MA\textsuperscript{a,b}
Allison B. McCoy, PhD\textsuperscript{d,e,f}
Scott Weingarten, MD\textsuperscript{g}
Robert Murphy, MD\textsuperscript{f}

Affiliations:
\textsuperscript{a} Partners Healthcare System, Boston, MA
\textsuperscript{b} Brigham & Women’s Hospital, Boston, MA
\textsuperscript{c} Harvard Medical School, Boston, MA
\textsuperscript{d} Department of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA
\textsuperscript{e} Center for Applied Health Services Research, Ochsner Health System, New Orleans, LA
\textsuperscript{f} The University of Texas School of Biomedical Informatics at Houston, Houston, TX
\textsuperscript{g} Cedars-Sinai Health System, Beverly Hills, CA

Abstract: (150-200 words)
Health IT (HIT), and clinical decision support (CDS) systems in particular, can improve the safety and effectiveness of health care when correctly implemented and effectively used. Over time, HIT and CDS have become more complex and pervasive, necessitating better monitoring of their efficacy and functionality. In some case, CDS alerts fire incorrectly or stop firing without warning, adversely affecting care quality and putting patients in danger. The issue now becomes, who shall watch the watchmen?

In this panel we present approaches to this issue from four sites’ perspectives, together with recommendations for best practices for monitoring CDS effectiveness. We also discuss how the value of an organization’s CDS can be measured.

The panelists will discuss the software and/or management plans they have developed in order to monitor CDS, the implementation of said software and plans, and the lessons learned. All four institutions represented on this panel are currently running self-developed CDS monitoring dashboards: Brigham and Women’s Hospital (Andersen), UT Physicians (McCoy), Cedars-Sinai Health System (Weingarten) and Memorial Hermann (Murphy) and have also formed CDS committees and governance approaches to oversee their large and diverse decision support implementations.

Description:
The overall outline of the talk is given in the abstract. The panel will be organized as follows:
<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Wright (moderator)</td>
<td>Introduction, overview of need for CDS monitoring and current approaches, through both software development and dedicated committees</td>
</tr>
<tr>
<td>15 min</td>
<td>Andersen</td>
<td>Predictive Analytics for CDS Effectiveness (PACE) dashboard</td>
</tr>
<tr>
<td>15 min</td>
<td>McCoy</td>
<td>InSPECt: Interactive Surveillance Portal for Evaluating Clinical support</td>
</tr>
<tr>
<td>15 min</td>
<td>Murphy</td>
<td>CDS Monthly Operating Report and CDS committee at Memorial Hermann</td>
</tr>
<tr>
<td>15 min</td>
<td>Weingarten</td>
<td>Cedars-Sinai CDS analytics dashboard</td>
</tr>
<tr>
<td>5 min</td>
<td>Wright</td>
<td>Summary of topics discussed</td>
</tr>
<tr>
<td>20 min</td>
<td>Panel</td>
<td>Discussion and questions from audience</td>
</tr>
</tbody>
</table>

The specific topics covered by each presenter will be:

1. **Wright (introduction):** Introduce the topic and panelists, including a brief description of what each panelist will be talking about and their affiliations. Provide an overview of the history of.

2. **Andersen:** Discuss the ongoing development and implementation of the Predictive Analytics for CDS Effectiveness (PACE) CDS alerts dashboard for Brigham and Women’s Hospital, which is focused on detecting malfunctions and anomalies in CDS [1]. Review the process of expanding the dashboard to other sites, feedback from test users, and current results.

3. **McCoy:** Present the development, multisite implementation, and evaluation of InSPECt (Interactive Surveillance Portal for Evaluating Clinical decision support), a web-based, Electronic Health Record (EHR)-independent dashboard for evaluating clinical decision support alert performance across sites with diverse healthcare settings, EHRs, and CDS configurations.

4. **Murphy:** Discuss the Memorial Hermann CDS Committee’s experience overseeing a large and diverse package of decision support, as well as their use of a Monthly Operating Report, focusing on alert performance using override rates and clinicians’ responses.

5. **Weingarten:** Discuss the development and use of the Cedars-Sinai CDS analytics dashboard. Review the Cedars approach to designing and measuring the financial impact of a package of CDS interventions based on the ABIM Choosing Wisely campaign.

6. **Wright:** Summarize topics discussed, identify common themes and outline the panel’s recommendations for best practices.

7. **Panel:** After the presentations, the panel will briefly discuss additional issues and then open for questions from and discussion with the audience. Audience members will be invited to address questions to specific panelists, or to the panel as a whole.

**Importance:**

Although the vast majority of hospitals and practices have now implemented some type of CDS (HealthIT.gov reported a 72.8% adoption rate of CDS in 2013), the ability to effectively monitor CDS alerts is currently lacking from most implementations. System upgrades, changes in rules, changes to external data sources such as a drug dictionary, or unintentional changes can cause alerts to spike, slow down, or completely stop firing, and without monitoring this can often go unnoticed for days, weeks, or as long as years. Consequences of these alert anomalies can range from mild (e.g. suggesting mammograms for a healthy 2-year old boy) to severe (e.g. an allergy alert failing to fire, resulting in a fatal allergic reaction).
Most hospitals and institutions completely lack the ability to monitor their CDS systems, and there is no standard tool to provide this functionality. A survey of chief medical information officers (CMIOs) found that 93% of CMIOs had experienced CDS malfunctions, and 62% were not confident that their current monitoring methods and tools could detect or prevent CDS malfunctions [1]. Despite this, the vast majority of medical institutions have no ongoing operational management plan for how to handle these malfunctions.

In order to monitor CDS, developers are often forced to rely on ad hoc reports which can be time-consuming to produce and intermittent. Consequently, it’s difficult to accurately assess if or how much CDS is adding value to the system as a whole, or to create a plan to optimize CDS. Furthermore, in HIT, there is often a disconnect between the developers implementing and maintaining the technology, and the clinicians using the technology. CDS data are usually stored on databases that require technical programming skills to access, creating a barrier to non-programmers who wish to monitor CDS alerts. Because there is currently no specific commercially-available software dedicated to CDS monitoring and no widely-accepted “best practices” for how to insure CDS is functioning correctly, many hospitals and institutions have chosen to develop their own software and/or management plans in order to address these issues, including CDS alert dashboards to access and display the data, and the formation of committees to oversee CDS implementation and functionality.

In this panel, we present the experience of CDS monitoring and evaluation from four different organizations. We focus on (1) detecting CDS malfunctions, (2) measuring CDS efficiency through alert acceptance rates and clinician feedback, and (3) evaluating CDS value in terms of clinical outcomes, safety, and cost savings.

We believe that monitoring and regularly assessing CDS is of critical importance to any organization that utilizes CDS. This panel will help to guide the discussion of CDS monitoring and work toward developing standards for CDS assessment that can be adopted by any organization with CDS.


Discussion Questions:
1. What tools and approaches can best help to detect CDS malfunctions and errors?
2. What are best practices for monitoring CDS and detecting and addressing CDS malfunctions?
3. How can organizations develop and employ effective approaches for CDS monitoring?
4. What are the limitations of CDS monitoring software?
5. What were the challenges encountered and lessons learned by the different organizations as they developed CDS monitoring methods and tools?
6. What are current unfilled needs and future potential for improvement in CDS monitoring practices?
7. How do you determine if your investment in CDS is yielding safety, quality and cost-effectiveness returns and which parts of your CDS are most valuable?
8. How can CDS monitoring improve CDS?

Participation statement:
All proposed panelists are aware of this panel submission, and have agreed to participate in the panel if the proposal is accepted.
PEDSnet: from building a high-quality CDRN to conducting science

L. Charles Bailey\textsuperscript{a}, MD, PhD, Michael G. Kahn\textsuperscript{b}, MD, PhD, Sara Deakyne\textsuperscript{b}, MS, Ritu Khare\textsuperscript{a}, PhD, Katherine Deans\textsuperscript{c}, MD

\textsuperscript{a}The Children’s Hospital of Philadelphia, Philadelphia, PA, \textsuperscript{b}Children’s Hospital Colorado, Aurora, CO, \textsuperscript{c}Nationwide Children’s Hospital, Columbus, OH

Abstract

Collaborations across multiple institutions are essential to achieve sufficient cohort sizes in clinical research. Within the national PCORnet research network, clinical data research networks (CDRNs) aggregate electronic health record (EHR) data from multiple sites and provide a unified platform to conduct large-scale studies, while dramatically increasing the capabilities to conduct a wide range of scientific studies. There are two critical facets of building a functional CDRN data resource: (i) to address the characteristic data provenance and quality issues that arise from the data’s orientation toward clinical operations rather than research, from semantic heterogeneity across systems, and from clinical data peculiarities, and (ii) to assess and improve the usability of network data to conduct scientifically valid studies. This panel focuses on a pediatric-specific CDRN, PEDSnet, to discuss its recent evolution from initial data collection and generic data quality assessments, to tailoring data characterization to scientific utility, developing a data-compatible approach to constructing computable phenotypes and conducting pilot observational studies.

Learning Objectives:

1. To understand the full CDRN cycle from validating network data quality to preparing tailored datasets for scientific studies
2. To learn about the iterative framework, guided by scientific requirements, for conducting data quality assessments and stabilizing network-wide ETL conventions in PEDSnet
3. To learn about challenges and opportunities in using the CDRN data for designing computable phenotypes and observational studies

Panel Description

PEDSnet, a pediatric learning health system, is a PCORI-supported CDRNs focused on child health. It aggregates EHR data from eight of the nation’s largest children’s hospitals. The digital infrastructure of PEDSnet has been iteratively developed through multiple data cycles with the participating sites over the last 2.5 years. PEDSnet has successfully addressed internal and external data quality requirements, and is conducting a number of observational studies as a test of network utility. In this panel, we invite leaders, developers and consumers of PEDSnet to discuss the journey of building, evaluating, and using this network for conducting scientific studies. The moderator leads the data coordinating center in PEDSnet, and the four panelists include (i) the PEDSnet informatics co-PI, summarizing the procedures for data specification and quality assessment in PEDSnet, (ii) a clinical informatician involved in the PEDSnet computable phenotype development program, (iii) a data scientist facilitating PEDSnet contribution to a PCORnet-wide study of medication exposure and child growth, and (iv) a clinical researcher, directing a study of diagnostic radiation dosage in collaboration with PEDSnet.

Moderator: Welcome and session overview
L. Charles Bailey, MD, PhD

Panelist #1: Promoting Data Quality Using Systematic Assessments and Structured ETL Conventions
Michael Kahn, MD, PhD (10 minutes followed by 5 minutes discussion)

Dr. Kahn will discuss the iterative framework of validating data quality in PEDSnet. The topics will include design of element-level data quality assessments, use of a structured document to validate and manage network wide ETL conventions, and external validation of the PEDSnet data against community standards and published results.
Panelist #2: Developing Computable Phenotypes in PEDSnet
Sara Deakyne, MS (10 minutes followed by 5 minutes discussion)

Ms. Deakyne will discuss her work on cohort identification in large-scale datasets. The topics will include definition and validation of computable phenotypes and case-finding algorithms in PEDSnet that could be replicated across other CDRNs.

Panelist #3: Designing Observational Studies on Antibiotics using PEDSnet
Ritu Khare, PhD (10 minutes followed by 5 minutes discussion)

Dr. Khare will discuss using PEDSnet as a testbed to support a PCORnet wide study of antibiotic exposure in infancy and subsequent obesity risk. The discussion will cover topics including characterization of exposures using clinical prescription, dispensing, and diagnosis data, and constructing growth assessments from primary data.

Panelist #4: Using PEDSnet data to conduct an observational study on radiation dosage
Katherine Deans, MD (10 minutes followed by 5 minutes discussion)

Dr. Deans will discuss her experience in designing a study of diagnostic radiation exposure using a combination of core PEDSnet data and study-specific radiation dosimetry. This talk will discuss the ways in which the study makes use of network data and collaborative infrastructure to answer specific research questions.

Moderator: Facilitated discussion and closing
L. Charles Bailey, MD, PhD (25 minutes)

Significance of Panel
PEDSnet has been recognized as one of the first functional CDRNs by PCORI. This panel will serve as a learning model for use of CDRN multi-site digital infrastructure to support large-scale scientific studies.

Discussion Questions
The following guiding questions will be used to facilitate discussion:
1. What are the key challenges in building a “high-quality” CDRN from informatics and science perspectives?
2. What unique opportunities does a CDRN offer investigators interested in conducting large-scale clinical research?
3. What are the main bottlenecks in conducting feasibility and scientific studies on network-based data?
4. What do we learn about data quality assessments based on the experience of computable phenotyping and designing observational studies?

Moderator Attestation
The four panelists, Dr. Kahn, Ms. Deakyne, Dr. Khare, and Dr. Deans, have all agreed to participate in this didactic panel.
From the Trenches- Issues Facing Clinical Informatics Administrative Clinicians in the Primary Care Setting

Deepti Panditia, MD\textsuperscript{1}, Curtis Boehm, MD\textsuperscript{1}, Jill Tiongco, MD\textsuperscript{2}, David Dorr, MD, MS\textsuperscript{4}
\textsuperscript{1}Park Nicollet-Health Partners Health System, Minneapolis, MN; \textsuperscript{2}Peninsula Primary Care, Carmel, CA; \textsuperscript{3}Oregon Health Sciences University, Portland,

Abstract
As primary care providers struggle to survive in an ever changing health care arena, 2016 is rapidly becoming the year of change. Primary Care Clinicians are increasingly taking on additional roles of Clinical Informatics Specialists and taking on challenging administrative roles along with Clinical Roles. This panel will illustrate some of the struggles faced by primary care providers in informatics related administrative roles such as CMIO and EMR champions, share examples of some of the major challenges and explore possible solutions that others can apply to their own health care setting. This will include strategies to manage change, meeting complex mandates, managing clinician expectations with changing work flows that impact and engage the entire care team and ultimately influence system design.

Introduction
As more and more Primary Care Clinicians are getting board certified in Clinical Informatics and taking on administrative roles in their health care institutions they feel that they are “in the trenches” without any support from rest of the administration due to the unique nature of their job. They also have to deal clinician opposition to change and inertia when it comes to embracing technology. This creates some unique challenges and clinicians who have done this for a while have valuable insight and possible solutions for these challenges.

Learning Objectives
1. Leadership and Governance strategies to improve adoption of EMR technology.
2. Understand barriers to Clinician acceptance to changes with HIT and EMR adoption
3. Understand complex workflow changes after EMR implementation and change management skills for managing these changes.
4. Challenges when transitioning from one EMR to another or merging EMR’s

Intended Audience
Any Clinician Informatics Specialist, Informatics and EMR Clinician Administrators, Primary Care and Specialty providers

Challenges facing CMIO and other similarly positioned Clinicians
Clinical leadership in healthcare informatics is an exercise in diplomacy, expectation setting and education. Dr. Boehm will address three topical areas of leadership in a large care delivery organization. Setting the stage and making the case for change. Evaluating the benefits of change for the organization, Tools of change management and adaptation.

Making the case for change: System changes should align with the organizations overall strategic plan. A key part of the CMIO role is to aid clinical leaders in understanding the capabilities of IT systems and partnership to develop a strategic IT plan.

Evaluation of systems to move from return on investment to organizational value: Traditional evaluation of system benefit has focused on return on investment. While this remains an important aspect of the evaluation needed, benefits accruing in clinical care include other dimensions that are not as easily quantified. Improved quality of care as reflected in quality metrics, reducing patient costs and improved patient and family experience are components that benefit the organization in less tangible ways much more difficult to quantify, such as patient loyalty and brand equity.
The importance of change adaptation strategies to achieve success: Clinical care teams live in an environment of constant change, awareness of the environment and a systematic approach to proactively create an environment supporting adoption increases the likelihood of achieving the goals of the project.

Understand barriers to Clinician acceptance to changes with HIT and EMR adoption

Crucial to any implementation of health systems is the ability to identify potential barriers to adoption. Dr. Tiangco will focus on what could hinder adoption to HIT and EMR changes in order to facilitate better implementation. Organizational experience, peer discussions and literature review have led to the proposal of two conceptual groupings of barriers to clinician acceptance. Identification of these type of barriers can alleviate the burden on users as it aims to identify "who can act on what." Numerous clinicians feel that barriers exist because of the intrinsic limitation of the health systems itself. Some however feel that barriers exist because of the limitations of what we, as clinicians could humanly do. The fact is, these are all true. Identifying the root cause of such barriers can therefore lead to overcoming them. So what are these barriers?

1. Static barriers: These are mostly finite resources. Finite resources include clinician’s time and money and vendor technical support. Various articles have made efforts to identify barriers and clearly, cost of implementation, decline in productivity during implementation, time required for data entry, EHR technical limitations and down times have been identified as the usual suspects. We then ask, can acknowledgement of these fixed and finite resources help modify and manage expectations and improve acceptance? How can stakeholders address these barriers?

2. Dynamic Barriers: These evolve and devolve. These include the need for constant workflow changes, policy changes, EHR transitions and changes, variances in metrics by payor etc. These barriers are so dynamic that the frequency of change itself leads to more frustration and lack of acceptance. The result is a constant yearning from clinicians for stability in the work environment to create efficiency. To attain stability is seemingly a unicorn at present day. Can clinician engagement facilitate acceptance?

Understand complex workflow changes after EMR implementation and change management skills for managing these changes.

Implementation of an EHR is only the start of the work for a clinical informatician in the trenches. In a seven year role as chair of the Clinical Decision Support Committee and as the leader of several innovative primary care programs for risk stratified care management, Dr. Dorr will discuss major issues around managing change to both workflow and future EHR “enhancements.” Specific discussion points will highlight: 1) the inter digitation of primary care workflow changes, especially population management, more care planning, and transitions of care, with the EHR changes made for implementation and needed for better alignment; 2) proactive management of EHR changes suggested by vendors and requested by users; 3) “anomaly” detection from EHR changes that disrupt workflows or break other functions in the EHR; and 4) the planning and implementation process for more complex endeavors. Specific topics covered will be primary care versus institutional governance of the process, encouraging learning both by induction and deduction, and overviews of the various workflows over time and how we broke and tried to fix them.

Challenges when transitioning from one EMR to another or merging EMR's

When transitioning from one EHR to another or transitioning for one version of the existing EMR to another there are several barriers experienced. Dr. Pandita will discuss the experience at Park Nicollet with their transition from legacy EHR to current EPIC EHR and now after we merged with Health Partners (HP) Health system we are in the process of transitioning Existing EPIC EHR to the HP Platform which brings its own set of complexities. The challenges with these changes are at several levels:

1. Technical challenges: Merging tools are based on classical logic and are forced to avoid inconsistencies which may cause valuable information loss which is an undesirable consequence as the lost information may serve as an
integral component in healthcare record. Avoiding redundancy in transition is another technical challenge. Backloading of data from legacy systems creates another interesting challenge particularly as it related to images, drug lists and problem lists.

2. User Acceptance: The changes in how clinicians and ancillary staff interact with “inbasket”, refills, phone notes, alerts and reminders and other messages can change with transitions and testing these ahead of time and including end users in training development is imperative.

3. Integration of legacy and new system with ancillary system vendors: This includes systems such as lab software, Cardiology and Radiology systems, ICU device integration and many more. Some face redundancy so might need to be abandoned and some need upgrades to be compatible with new EHR.

**Statement of Agreement and Support**
All members of this panel are in agreement to participate in this discussion. This panel is endorsed by the Primary Care Informatics Working Group.
VISUALIZATION OF PATIENT-REPORTED OUTCOMES

Adriana Arcia, PhD, RN¹, Kenrick Cato, PhD, RN ¹², Ruth Masterson Creber, PhD, RN¹, Yalini Senathirajah, PhD³ and Sunmoo Yoon PhD, RN¹

Abstract
Patient-reported outcomes (PROs) are outcomes that patients self-report either through self-administered questionnaires or interviews. Examples of PROs include symptoms, functional status, and health-related quality-of-life. While visualizations of PROs can facilitate communication between client and clinician and inform patients’ efforts to self-manage health, the visualization creation process is often arduous.

This panel is intended to provide researchers, clinicians, educators, and policymakers with an introduction to methods for development and testing of visualizations of PROs. To this end, each panelist will discuss issues specific to each phase of visualization development for PROs (design, automation/programming, testing, and selection/deployment) with a particular focus on those generated by the Patient Reported Outcome Measurement System (PROMIS).

Learning objectives.
After participating in this session, the learner should be better able to:

- Provide rationales for the use of visualization.
- Describe the phases of the visualization process and the key activities of each phase.
- Understand the triangulation approach to selecting visualizations based on qualitative and quantitative methods.

Panel Description
The first presentation will focus on the Design Phase of the visualization process. Case studies will serve to illustrate key activities of the design phase including: defining desired outcomes, matching data types to graphical formats, drawing on culturally appropriate graphical conventions, and iterative prototyping with feedback from the intended audience. The case studies feature tailored infographics of self-reported health data, including PROMIS measures, from Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) and the New York City Hispanic Dementia Caregiver Research Program.

The second presentation will focus on automation and programming using a mental health example. The first step is to collect data from non-expert patients about their health experiences and mental states, using technologies such as tablets which must be appropriate to the language and levels, technology knowledge, privacy concerns and other patient characteristics, often at home or in a clinic setting. The second
step requires the clinician to perform close management and rapid interpretation for making a diagnosis or for treatment monitoring, over time as well as in single instances. Patients and providers are limited by time for learning so the design must be appropriately simple and either convey affordances quickly for one-shot experience or leverage known conventions. The same is true for researcher stakeholders.

The purpose of the third presentation is to demonstrate how tailored infographics, based on a comprehensive symptom assessment, can serve to influence a patient’s mental model of their symptoms. Using qualitative interviews with both patients and providers, we will test whether specific visualizations from a novel mHealth app can effectively represent patients’ symptom burden. We will also test whether or not they are useful tools for effectively communicating symptoms between patients and providers.

The last presentation focuses on the selection and deployment of visualizations after applying triangulation approaches following multiple evaluations. This section will walk through a triangulation strategy to choose the right visualizations from qualitative (e.g., interview, focus group, think-aloud) and quantitative (e.g., eye tracking, comprehension test) studies with the following three multi-method visualization projects; 1) visualizations for clinicians in an acute-care setting, 2) visualizations for patients in a community setting, and 3) visualizations for health researchers in a research institute setting.

**Questions**

- What infrastructure and cooperative agreements would be required to set up a bank of visualizations of PROMIS and other self/patient reported outcomes?
- A systematic approach to visualization is a substantial undertaking. When is it worth the effort?
- How do factors such as language, culture, and health literacy affect visualizations’ effectiveness?
- Can we make a decision on choosing visualizations based on focus groups?
- What are the implementation barriers perceived by researchers and how could implementers ease this burden?

1. Columbia University School of Nursing, New York, NY
2. New York Presbyterian Hospital, New York, NY
3. Dept. of Medical Informatics, SUNY Downstate Medical Center, Brooklyn, NY
The OpenMRS Community’s Experience: A Decade of Developing and Implementing Medical Record Systems Within Resource Constrained Countries

Theresa Cullen, MD, MS1,2, Paul Biondich, MD, MS1,2, Burke Mamlin, MD, MS1,2, Judy Wawira, MD, MS2, Hamish Fraser, MRCP3
1Global Health Informatics, Regenstrief Institute, Indianapolis, IN; 2Indiana University School of Medicine, Indianapolis, IN; 3Leeds Institute of Health Sciences and the Yorkshire Centre for Health Informatics, Leeds, United Kingdom

Abstract
In 2016, the Open Medical Record System (OpenMRS) community celebrates its ten year anniversary. OpenMRS, a collaboratively developed free and open platform for building electronic medical records (EMR), is currently in use in over 1200 facilities, and supports care provision to over 5 million patients [1]. These software components are developed and supported through the OpenMRS rapidly evolving and engaged multinational open source community.

This panel is designed to share the experience of the last ten years from the vantage point of key community participants who have helped lead the community over the last decade. The panel will discuss the history of OpenMRS, current state (including a technical overview and description of the close alignment with standards development activities), implementation challenges faced, and future growth plans. The panel presentations will emphasize lessons learned, including predictive factors for a successful open source health information technology community [2].

Introduction and Background
The benefits of open, collaboratively developed health information technologies (HIT) have been well described, and strongly advocated for over the past decade [3]. However, most have found it difficult to successfully instigate, grow, and maintain these initiatives. As the number of sustainable open-source HIT solutions appear to wither within the US, the ongoing experiences within the rapidly growing OpenMRS community provide a potent reminder of the possibilities and promise of open source HIT.

While the origins of OpenMRS have their roots in work that started in 2004, the initiative is currently celebrating its tenth anniversary as an open source community. OpenMRS represents a successful collaborative action network, where many disparate organizational and individual collaborators commit to a shared mission of developing and implementing EMRs for underserved populations, and work together on and through shared governance, technologies, standards, principles, and resource gathering.

Panel Description
Dr. Terry Cullen will lead a panel discussion on the history and current status of OpenMRS, and in doing so will help the audience better understand the importance of open source communities, the many collateral benefits that emerge from them, and the potential for using the OpenMRS experience to help inform current US based healthcare conundrums. The panel will include a review of various implementation activities and the factors that contributed to their success. The panel will also discuss the historical and future management strategies employed by project leadership, which will provide details on the nuanced decisionmaking that proved fundamental to the initiative’s
success. The experiences and lessons learned from this global health informatics initiative are likely applicable to not only HIT development and implementation initiatives working under fiscal and human constraints, but to the US situation as well. Given that AMIA strives to be a professional home for global health informatics activities, it is well suited for this conference.

**Presenters**

**History of OpenMRS (Fraser)**

The panel will begin with Dr. Hamish Fraser’s overview of the early history of the initiative. Dr. Fraser is currently the Associate Professor of eHealth at the University of Leeds, and was one of the original founders of the OpenMRS community. He will briefly describe the original collaboration between two global health informatics groups both seeking to support EMR implementations within two diverse clinical care settings in sub-Saharan Africa. He will describe how this work evolved into a global consortium of hundreds of developers, implementers and end users of e-Health technology focused on improving the health of and empowering the underserved. This will include a review of the mission, vision, and values of the community, and a description of the community processes designed to encourage broad participation of individuals and organizations within the OpenMRS community. This presentation is designed to familiarize the audience with the project and begins to help elucidate the factors (such as a shared mission, vision and values) that are foundational to the community’s continued growth and success. These factors have meaning within the global community at large.

**Technical Overview and Development Community (Mamlin)**

Dr. Burke Mamlin is an associate professor of medicine and the Chief Architect of OpenMRS. In his presentation, Dr. Mamlin will describe the technical underpinnings of OpenMRS, including its data model, its application programming interface (API), and flexible user interface framework. In describing the specifics behind the modular, extensible platform, he intends to demonstrate how OpenMRS can be and has been so broadly implemented in diverse clinical settings. His presentation will demonstrate how OpenMRS has benefited from and contributed to standards over the past decade, including work on interoperability. He will describe how the OpenMRS Developers Community is itself being developed, undergoing ongoing continuing improvement and finding innovative ways to motivate and recognize developers in the community. Finally, he will discuss how OpenMRS is leveraging many of the latest technological advances including RESTful web services, support for FHIR, and integration with Observation Health Data Sciences and Informatics (OHDSI) community [4]. The prescience of OpenMRS to develop and support a modular and flexible user framework should help inform the next generation of US HIT systems.

**Implementation (Gichoya)**

Dr. Judy Gichoya is currently a third year radiology resident at Indiana University, where she previously completed a health informatics fellowship. She has been active in the OpenMRS community for the last decade, and has been responsible for several successful OpenMRS implementations. Today, OpenMRS is implemented in well over 1200 facilities. Evaluation of these implementations has shown that there are key factors to successful implementation, independent of location. These factors include community involvement and ownership, support from the Ministry of Health or appropriate local organizations, and a commitment to local empowerment [5] [6]. Dr. Gichoya will give a presentation on the factors that contribute to successful implementations of medical record systems such as OpenMRS within constrained environments. She will illustrate this through sharing her personal experiences implementing OpenMRS in health care facilities in downtown Nairobi, and through highlighting other representative implementation examples around the world. Additionally, Dr. Gichoya will share results of the qualitative work she is currently performing within the OpenMRS community specifically focused upon critical implementation success factors. This presentation will once again illustrate the consistency of factors that contribute to implementation success independent of locale.
Future Directions and Lessons Learned (Biondich)

Dr. Paul Biondich is one of the founders of OpenMRS and serves as the executive project lead for the OpenMRS community. OpenMRS now serves as a foundational technology for a half dozen national-level EMR implementation activities in countries both in Africa and Southeast Asia. This substantive commitment to the platform and the community provoked a recent strong emphasis upon strategic planning and community governance. Large scale implementation activities also necessitate the development of regional business ecosystems and substantive international workforce training programs. Dr. Biondich will review the OpenMRS 2016 Community Strategic Plan [7][8], and discuss the approaches designed to respond to current demands and their rationales based upon previous experiences. Dr. Biondich will additionally synthesize overarching lessons learned both from the previous panel presentations that can help address the HIT questions facing the US health care system such as care coordination.

Discussion Questions

1. What are the top three benefits and challenges of open source development for global health IT?
2. Is there a role for AMIA in helping identify and sustain open source HIT communities, either within the US or globally?
3. How can AMIA leverage the lessons learned from global HIT development, like OpenMRS, to benefit the US HIT and health informatics community?
4. What are appropriate licensing options to support community growth in open source HIT?
5. What can AMIA do to assist the global health informatics community?

Panel Organizer Statement: All participants have agreed to take part in the panel and discuss the topics as outlined above.

References

8. OpenMRS Operational Plan available from: https://docs.google.com/document/d/10TO1iG-W05pblq36VW_UDVK6uNHi4KKEFphAfIVunw/edit?usp=sharing
Didactic Panel: **Watson Cognitive Computing for Electronic Medical Records**

Moderator:

Murthy V. Devarakonda, PhD, IBM Research, Yorktown Heights, NY

Panel Members:

1. Neil Mehta, MBBS, MS, Cleveland Clinic, Cleveland, OH
2. Christopher Nielson, MD MPH, VHA, Reno, NV
3. Kenney Ng, PhD, IBM Research, Yorktown Heights, NY
4. (Ms.) Preethi Raghavan, PhD, IBM Research, Yorktown Heights, NY

**Abstract**

With the explosive growth of data, the human brain is substantially overloaded with more information than it can make sense of. This is painfully clear in patient care where critical information may be buried in the mountains of data entered into the Electronic Medical Record (EMR) system while clinicians struggle to make sense of this information at the point of care. Cognitive computing, exemplified by Watson, offers the promise to transform this data to an intelligent assistant that provides precise insights when needed. This didactic panel examines recent advances in the specific capabilities of such cognitive computing and prospective benefits of these capabilities in patient care. The panel will discuss cognitive needs of practicing physicians. Watson capabilities of automatically summarizing a patient record and natural language question-answering on a patient record as examples of intelligent assistance cognitive computing can provide. The panel will also discuss complementary technologies from massively large scale data analysis and knowledge management that will play a role in cognitive computing. The learning objectives include developing an appreciation for working examples of cognitive computing for Electronic Medical Records, and an understanding of how these capabilities could mitigate physician’s cognitive load.

**Description of the Panel and Issues Examined**

The panel consists of two practicing physicians, Drs. Neil Mehta of Cleveland Clinic and Christopher Nielson the Veterans Health Administration, and two highly experienced IBM Watson technology researchers, Kenney Ng, PhD, and (Ms.) Preethi Raghavan, PhD. Dr. Mehta has additional research interests in technology enhanced teaching and medical informatics. Dr. Nielson has deep experience in big data and predictive analytics. Mr. Ng is a research scientist and a technical manager at IBM research specializing in patient similarity and predictive analytics. Ms. Raghavan is a research scientist in IBM Watson Health and specializes in discovery of semantic structure in natural language text using statistical machine learning and its application to medical informatics.

The panel will examine the potential of Watson and Watson-like technologies to meet the information needs of physicians in patient care, in the context of Electronic Medical Records. IBM Watson has unveiled an era of cognitive computing, and a broad set of applications are possible using the technologies. One particular area of immediate and urgent need is to apply the technology to address the cognitive load on physicians in using the patient data in Electronic Medical Records. To be effective in this context, it is necessary for the technology to support the workflow and the information requirements within the workflow. Prior to a patient visit, a physician may be seeking information such as, “What was done at the last visit? What data has accumulated since the last visit? What is overdue or needs to be addressed today?” without being overwhelmed with irrelevant information. Furthermore, it is also important to understand pain-points with the state of the art.

**Dr. Neil Mehta** will open the discussion with a practicing physician’s perspective of the informational challenges in using EMRs. It is well documented that there is a low physician satisfaction with most EMR systems as they fail to live up to their potential. Physicians and other providers spend a lot of time and effort entering data into the EMR (often following exhaustive institutional requirements of structured input and completeness) but do not see
tangible results commensurate with that investment and with their expectations created by the other internet
technologies. As the amount of data in the EMR increases, it takes more time and effort to get back useful
information from the EMR that would help with diagnosis and decision making. Caregivers are well-aware of
systems that have revolutionized other industries and are frustrated with the slow pace of change. A well-
designed effective cognitive computing system can play a huge role in solving this problem by giving valuable
“smart” information to the caregiver based on information in the EMR and evidence in the literature. Such
contextualized evidence based information would help EMRs live up to their potential and bring out a positive
change in healthcare of patients and populations.

Building an effective cognitive system takes many components. IBM Watson technology has demonstrated the
ability to analyze large volumes of medical literature. More recently IBM Watson has evolved to analyze the clinical
notes and structured data of a patient record into a meaningful summary of the record, centered around
automatically generated problem list. In order to build these EMR-centric analysis and even more advanced
decision support, millions of existing records can be an invaluable source of knowledge complementing the
knowledge from the medical literature.

Dr. Christopher Nielson will provide another perspective of the effects of increasing amount of data in patient
records, and in addition he will focus on how millions of such records can be analyzed to help develop knowledge
that enables cognitive computing. As increasing amounts of electronic data becomes available, fundamental
relationships that are critical to diagnosis and medical management may be obscured by irrelevant data. Although
the need for optimized data presentation is well recognized, manual identification of important relationships
between thousands of data types is impractical. The VA EMR database includes 30 million records with 15 years of
continuous data including nearly all data domains. The availability of such a large database permits probabilistic
evaluation of relationships between data domains with respect to needed associations, including frequency,
specificity and temporality. Rather than providing data simply organized by alphabetical order or location,
information can be presented with the patterns that are critical to diagnosis and management. Probabilistic
relationships create a foundation for decision support as well as a problem oriented approach to medical
management.

Kenney Ng will explore the power and promise of predictive modeling and patient similarity analytics with
concrete examples of how these methods can be applied to improve patient care. Experience shows that the
predictive models can generate actionable insights in addition to the risk score. In addition to identifying the top
risk factors, methods can also identify patient specific, time dependent risk factors. Patient similarity analysis, a
complement to the predictive modeling, is an approach to support personalized medicine. In this approach, a
similar patient cohort is first determined for a given individual patient and, machine learning and data mining
techniques can then be applied to the similar patient cohort to derive actionable insights. Customized models can
identify patient specific risk factors instead of just population level risk factors. Another application is to combine
patient similarity with drug similarity to tailor treatments to individual patients based on their likelihood to
respond to a therapy. The discussion will present results from studies on patient data.

Preethi Raghavan will discuss the potential and the promise of natural language question answering (QA) based
approach to managing the cognitive load in accessing EMR content. There is a wealth of patient information buried
in EMRs even as they are being recorded using today’s EHR systems, and we want to enable efficient discovery of
that information. The challenges include interpreting and making inferences from the structured and unstructured
text (without imposing additional requirements on the data entered) in the EMR to provide physicians with the
required answers to support clinical decision-making. Multiple strategies are being explored to navigate the vast
amounts of information about a patient. Some of these strategies, such as proactively summarizing important
patient information and a concept-based search in clinical notes and structured data, have been prototyped in IBM
Watson as well as in other research projects with encouraging results. However, the ability to precisely answer
patient-specific questions at the point of care would be even more effective. For example, “Were there any
complications of the patient’s RYGB surgery?” could be answered with very specific phrases, passages and
structured entries from the patient record. The discussion will examine different challenges associated with question answering on EMRs when compared to open-domain factoid QA, and how advanced natural language processing and patient modeling techniques can address these new challenges in QA on an EMR.

**Demonstration of IBM Watson:** We will demonstrate the present prototype IBM Watson patient record summarization and semantic search capability as an example of the cognitive computing discussed here.

**Relevance of the Panel and Anticipated Audience**
Cognitive load in accessing patient-specific information an EMR is a clear and present challenge. Physicians are frustrated with the current EMR technology. Cognitive computing exemplified by IBM Watson appears to have the potential to alleviate the cognitive load. A discussion of how these two trends intersect with a positive outcome is educational and deserves a detailed discussion. It is particularly interesting when the panel involves clinical informatics researchers and patient care practitioners on the same stage. Because of the panelists experience and their current research, it becomes possible to see how these possibilities go beyond just a hope and translate into results with real data. Because of the significance of these results in patient outcomes, this panel would be a significant attention grabber.

The audience for the panel will be from a very broad spectrum of the AMIA membership. Members with medical background, practicing or otherwise, would be interested to learn the cognitive computing capabilities that are feasible and can potentially help in patient care. Members without the medical background but with statistical and computer science background learn the cognitive needs of physicians and the state of the art for the cognitive computing analytics. Students, post-docs, and other early career researchers learn realistic challenges in patient care and how concrete examples of cognitive computing applications.

**Discussion Questions**
1. What is the precise nature of the cognitive needs of a practicing physician?
2. Why is there a cognitive needs gap?
3. How can big data analytics of millions of existing patient records help cognitive computing?
4. What is the current status of IBM Watson-like cognitive computing in the context of EMRs?
5. What are the results that indicate cognitive computing helps physicians?
6. How can predictive analytics be customized for individual patients?
7. How can patient similarities help derive actionable insights and tailored treatments?
8. What are the challenges of natural language question answering on an EMR?
9. What are the promising solutions for developing natural language question answering on an EMR?

**Statement of Participants’ Agreement**
I received positive acknowledgement in the form of email messages from the panel participants agreeing to participate in the panel.
Wait, my patient is where? Promises, challenges, and impact of automated event notification systems

Panelists
Brian E. Dixon, PhD, MPA, FHMSS\textsuperscript{1,2,3}, Jason S. Shapiro, MD, MA\textsuperscript{4}, Jessica S. Ancker, PhD, MPH\textsuperscript{5}, Joshua R Vest, PhD, MPH\textsuperscript{1,2}

Moderator
Saira N. Haque, PhD, MHSA\textsuperscript{6}

\textsuperscript{1}Regenstrief Institute, Indianapolis, IN; \textsuperscript{2}Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; \textsuperscript{3}Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service; \textsuperscript{4}The Icahn School of Medicine at Mount Sinai, New York, NY, \textsuperscript{5}Weill Cornell Medical College, New York, NY. \textsuperscript{6}RTI International, Research Triangle Park, NC.

Abstract
Electronic event notifications, also called alert or subscription services, provide real-time automatic alerting of patient health care encounters to healthcare providers at other institutions. For example, an ambulatory primary care provider might use the service to sign up for alerts when his/her patients are admitted to regional emergency departments or inpatient departments. In this didactic panel, panelists will discuss use cases for event notification systems, the current state of the evidence about whether these systems affect healthcare quality or cost, sociotechnical considerations in developing these systems, and ongoing evaluation work to move the evidence forward.

Keywords
Health information exchange; workflow; organization and administration; evaluation

Introduction
Event notification is the real-time, electronic, automatic alerting of providers to their patients’ contact with other health care facilities\textsuperscript{(1)}. Also referred to as an alert or subscription service, event notifications are messages triggered by specific patient activities, such as visit being registered in a hospital's admission-discharge-transfer (ADT) system. Triggers often involve an inpatient admission, discharge, or emergency department visit. Event notification systems fall under the broader category of health information exchange (HIE) technologies because electronic patient information is being shared between different organizations\textsuperscript{(1)}.

Event notification may address some of the information and communication challenges created by a fragmented health care system\textsuperscript{(2)}. A typical use case for event notification is as mechanism to inform an ambulatory primary care provider in near real-time that one of his or her patients has been admitted to a local emergency department or inpatient facility\textsuperscript{(3)}. In some instances, the event notification message be routed directly to a physician. More commonly, health care organizations receive event notifications and rely on members of a health care team, such as a case manager, care coach, or patient navigator to receive, manage and respond to alerts\textsuperscript{(4)}.

Event notifications offer several potential opportunities for improving health care quality. Event notifications are a new source of information about patients, thereby increasing providers’ general awareness of patients’ complex medical histories\textsuperscript{(5,6)}. Additionally, informing the provider by event notification creates opportunities for immediate intervention where appropriate\textsuperscript{(1)}. For example, health care providers may be able to contact the emergency department prior to inpatient admission to share information about the patient. Likewise, if the patient is still at the emergency department or hospital, the ambulatory primary care provider could support the coordination of post-discharge transitions and services, potentially improving the quality of care for a patient\textsuperscript{(3)}. Also, event notification may help identify patients for referral into care coordination programs\textsuperscript{(7)}. Alternatively, it is possible to structure event notification to identify missed health care contacts like a no show for specialist appointment or failure to obtain a prescription refill.
Event notifications appear to be a popular service as they have quickly been adopted by various HIE initiatives around the U.S. Additionally, they theoretically support the population health aims of integrated delivery systems(3). The evidence base for the effectiveness of event notifications is just beginning to grow. Few studies have examined these systems in real-world settings.

The panel

This panel seeks to summarize the challenges associated with event notification systems and the potential impact of these systems on care and decision making processes. Questions the panel will address are:

1. What are the use cases for event notification systems?
2. What is the role of organizational capacity and workflow fit in the implementation of event notification systems?
3. What are the challenges associated with effective implementation and use of event notification systems?
4. What is the relationship between event notification systems and other forms of health information exchange such as query-based exchange or DIRECT secure messaging?
5. What is the evidence that event notification systems can affect care delivery?

Panelists

The panelists have researched and evaluated event notification systems in diverse settings, with differential evaluation approaches, and from different organizational and theoretical perspectives.

Dr. Brian Dixon is an informatics researcher at both the Regenstrief Institute and the U.S. Department of Veterans Affairs. Dr. Dixon has worked with the Indiana Health Information Exchange to implement event notification services to private health system customers. Furthermore, he is currently the Indianapolis site lead for a new grant within the VA that is evaluating event notification services for older Veterans at two VA medical centers. In the VA project, Dr. Dixon is establishing triggers for when older adults seek care outside the VA. The notification messages will notify the Veteran’s primary care provider and, for those enrolled in the intervention arm of the study, a care coordinator in the Geriatrics service within VA. The care coordinator will work with the patient’s inpatient or ED facility to coordinate a discharge back to home where the VA can then provide intensive home-based care plus scheduling the patient for a follow up visit with his or her primary care provider. Dr. Dixon will describe the study design and early lessons to date rolling out the intervention across the Indianapolis VA region.

Dr. Jason Shapiro will discuss the differences between subscription-based event notification services and analytics-based notification services. Subscription-based event notification services usually require that a list of patients on whom notifications are desired be generated by the provider organization and sent to the health information exchange where surveillance is then done for trigger events. Analytics-based event notification services use a set of pre-determined criteria to define a cohort of patients on whom notifications should be sent using specific rules for trigger events. Examples of each will be given, and potential future use cases will be discussed.

Dr. Jessica Ancker will describe a survey of health care professionals who receive event notifications as part of their daily practice or job. The research, adopting a sociotechnical perspective, sought to identify the role of organizational features, workplace policies and procedures, and individual attitudes on perceptions of event notification quality, efficiency and satisfaction. Findings suggest that health care organizations with appropriate processes, workflows, and staff roles may be better positioned to use event notifications. Additionally, information quality remains critical in users’ assessments and perceptions. The work also identified specific instances where providers reported event notifications changed medical decision making and clinical activities.

Dr. Joshua Vest will describe the findings of a quantitative assessment of the impact of event notification on hospital readmission rates. Dr. Vest and his colleagues examined the event notifications offered in the Bronx, NY in a cohort of Medicare beneficiaries over a period of 54 months. Their evaluation compared the demographics and risk of the beneficiaries subscribed to event notifications and those beneficiaries who were not subscribed. Additionally, they found event notifications were associated with lower readmission rates.

As Moderator, Dr. Saira Haque will introduce health information exchange and event notification. She also will identify common themes across Drs. Dixon, Shapiro Ancker and Vest’s presentations. Dr. Haque will also review cross-cutting facilitators, barriers and considerations with event notification across sites.
**Learning objectives**

After participating in this session, the attendant will be able to:

- Understand the goals and basic functioning of an event notification system.
- Identify the challenges associated with event notification systems.
- Suggest methods to improve the fit of event notification systems into ambulatory care practice.
- Describe the evidence-base on the effectiveness of event notification systems.

This panel has been endorsed by both the Evaluation (Eval) and the People and Organizational Issues (POI) working groups.

**References**

Ensuring Reproducibility in Observational Research: 
Building and Sharing Knowledge Resources in the OHDSI Network

Jon D. Duke, MD MS¹, George Hripcsak, MD MS², 
Patrick Ryan PhD³, Nigam Shah MBBS, PhD⁴
¹Regenstrief Institute, Indianapolis, IN; ²Columbia University, New York, NY; ³Janssen Research and Development, Titusville, NJ; ⁴Stanford University, Palo Alto, CA

Abstract
The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) was formed in 2013 with the goal of creating reliable scientific evidence through large-scale analysis of observational health data from around the world. To ensure transparency and reproducibility in our research, we have created tools to share knowledge artifacts and other resources efficiently amongst OHDSI community members. We have also worked to facilitate best practices in the validation and governance of these resources. In the proposed panel, we will discuss our approach to building shareable knowledge assets, conducting research using these assets, and supporting access to these resources beyond the OHDSI community.

Organizer: Jon D. Duke, MD, MS

Background and Panel Description
The Observational Health Data Sciences and Informatics collaborative (OHDSI, pronounced ‘Odyssey’) was formed in 2013 with the goal of creating reliable scientific evidence through large-scale analysis of observational health data from around the world. To advance this goal, OHDSI has had to rapidly expand its scientific, technical, and community infrastructure. OHDSI has now grown to over 140 participants from 16 countries. Across the collaborative, there are 84 databases covering over 600 million patient lives that have been transformed into a common data model. OHDSI has conducted multiple international network-based observational research studies using this infrastructure. In this panel, we focus on resource sharing in the OHDSI community. Specifically we will discuss how knowledge assets such as clinical phenotypes can be consistently generated for reuse and consumption by a variety of stakeholders. By the end of this panel, attendees will gain an understanding of the issues and approaches to resource sharing in large-scale observational research.

Panel Importance and Target Audience
The proposed panel demonstrates how international collaborative research can be conducted in a transparent and reproducible fashion through shared resources. The availability of common knowledge artifacts and tools to create and share such resources addresses a critical need for researchers and supports the recent NIH guidelines on transparency and reproducibility in biomedical research.¹ This panel will provide attendees with insight into the opportunities and challenges of conducting reproducible research across diverse datasets as well as a clear path to utilizing OHDSI resources to support reproducible research, even outside the OHDSI network. The target audience includes health informaticians, clinical and health services researchers, health IT leaders (e.g., CMIOs), implementation scientists, epidemiologists, biostatisticians, and data scientists.

Presentations

Enabling Reproducible Research from Concept to Publication (P. Ryan)
Reproducibility - the ability to duplicate results from an experimental procedure - is a cornerstone of the scientific method. Analyses of observational health data have been challenged due to their apparent lack of reproducibility, but different researchers with access to the same data or researchers applying the same methods to different data. In
In this context, the problem can be deconstructed into two challenges: 1) the lack of standardized processes that can be transparently and consistently applied to a given dataset to replicate an analytical result; and 2) inconsistent generalizability across populations that can arise from patient-level heterogeneity. The OHDSI community is working together to establish consensus scientific best practices for clinical characterization, population-level estimation, and patient-level prediction, and is embedding these best practices through open-source software that enable standardized analytics that generate real-world evidence is a reproducible fashion. We will argue that through standardizing the design, execution, and reporting to community-consensus and empirically-based best practice, we can simultaneously improve the efficiency, accuracy, and reproducibility of observational analyses.

**A Shared Library of Validated Computable Phenotypes for Observational Research (J. Duke)**

Phenotype definitions are a central component of observational research. Patient cohorts must be transparently and explicitly defined to allow others to interpret and replicate a research study. In the OHDSI community, we are seeking to facilitate best practices in development, validation, and sharing of patient phenotypes through the establishment of an OHDSI Cohort Definition Library. All cohort definitions in the OHDSI Library should include 1) a human readable cohort definition listing all inclusion rules, exclusion rules, and referenced concept codes; 2) a machine-computable JSON syntax explicitly representing the above definition; 3) a validation study configuration to support consistent evaluation of cohort definition performance (e.g., precision, recall) across multiple datasets; 4) additional metadata on provenance, supporting literature, and performance results where available. It is hoped that by incorporating computable definitions of not only the phenotypes but of their validation criteria as well, we will encourage greater consistency in constructing and validating cohorts for collaborative research.

**Building Integrated Knowledge Resources for Drug Safety (N. Shah)**

Drugs are among the most common interventions examined via observational studies. Given the volume of research activity in drug-safety research, and the widespread use of spontaneous reporting systems (SRS) data for hypothesis generation and validation of newer approaches, there is an urgent need to create integrated knowledge sources that support evaluation of drug safety surveillance methods. We will describe a freely available, curated and standardized version of the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) data, which we call AEOLUS for Adverse Event Open Learning through Universal Standardization. In the creation of this public resource, as part of our work in the OHDSI community, we remove duplicate case records, apply standardized vocabularies to source codes and terms, mapping drug names to RxNorm concepts and outcomes to SNOMED-CT concepts, and pre-compute summary statistics about drug-outcome relationships. We make this resource, along with the source code to refresh the dataset, freely available to reduce the amount of time spent performing data management on the source FAERS reports, improve the quality of the underlying data, and enable standardized analyses using common vocabularies across disparate evidence sources.

**Embedding Large Scale Analytics at the Point of Care (G. Hripcsak)**

Modern research and modern medicine require access to information. Institutions sometimes struggle with providing clinical data to its research and clinical staff, usually offering a clinical data warehouse and a finite number of analysts to answer questions, generally after an approval process that may involve an institutional review board and other institutional committees. To facilitate information access, Columbia offers to its 12,000 research and clinical staff immediate access to an OHDSI database using the OHDSI Atlas tool set to query and analyze data. The data are de-identified, and IRB permission has been obtained on behalf of staff to query the database directly and immediately. The Atlas tool set helps users to navigate the terminologies, assemble correct queries, and visualize the results.

**Discussion Questions**

- What are the main challenges associated with performing reproducible research?
- What is the value of computable knowledge assets such as clinical phenotypes?
- What is gained and what is lost in the process of creating a computable phenotype?
- How can phenotype performance be consistently evaluated across diverse institutions?
- What is the difference between coded data and standardized data?
- How can the quality of a transformed data set be evaluated and, where necessary, improved?
- What are the barriers to sharing aggregate clinical data to providers within an institution?
- How can the interpretation of aggregate clinical data be facilitated by common reporting libraries?

Assurance
The organizer Jon D. Duke vouches that all listed participants have agreed to take part on the panel.

References

The Best of Imaging Informatics Research 2016

Panelists: Charles E. Kahn, Jr., MD, MS¹, William Hsu, PhD²
¹Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA; ²Department of Radiological Sciences, University of California, Los Angeles, CA

Abstract

The field of imaging informatics is rapidly advancing in its ability to address challenges related to clinical big data and harnessing this information for precision medicine. In the past year, the field has experienced growth in a variety of areas including radiomics (the generation of high dimensional features from images), development of new ontologies and standards for capturing information from images and reports, and unsupervised learning from images to predict the course of a disease and treatment response. In addition, we have seen a remarkable growth in novel approaches that go beyond pixel data by integrating imaging with other biomedical data, standardizing imaging workflows, and improving the quality and utility of image-derived information in clinical practice. In this session, we will review key advances in imaging informatics research published this past year.

Overview

Biomedical informatics spans a spectrum from basic to applied research, and from molecular to population scale. Imaging informatics is a discipline that focuses on improving patient outcomes through the effective use of images and imaging-derived information in research and clinical care. In the past, imaging informatics at AMIA has focused primarily on management, processing, and analysis of imaging data. As the storage and display of multidimensional image data has become commonplace, imaging informatics has evolved from primarily characterizing information from pixels and voxels in images to integrating evidence from images and other biomedical data sources to provide a systems-level understanding of disease etiology and progression. These developments have underscored the need to engage individuals from different disciplines to address shared challenges in managing heterogeneous datasets, developing common standards, and interpreting multiscale evidence for precision medicine. This session provides attendees with a topical review of new and interesting developments in imaging informatics in areas such as health policy, ontologies, standards and interoperability, reporting, radiogenomics, and deep learning. We expect this session to appeal to AMIA attendees given its emphasis on shared challenges and techniques such as data standards, information extraction, and predictive modeling that are broadly studied and applicable across the biomedical informatics field.

Methods

Searches using PubMed and Google Scholar will be performed relevant imaging informatics keywords and filtered for the previous twelve months. In addition, the abstracts of articles published in publications such as Radiology, RadioGraphics, Journal of the American Medical Informatics Association, Journal of Biomedical Informatics, and Journal of Digital Imaging, arXiv, Journal of Pathology Informatics, Neuroimage, Medical Image Analysis, and Academic Radiology will be reviewed. These searches will be supplemented by a call for nominations from luminary imaging informatics groups, including the Biomedical Imaging Informatics Working Group of AMIA, the Radiology Informatics Committee of the Radiological Society of North America (RSNA), the Informatics Commission of the American College of Radiology, and the Board of Directors of the Society for Imaging Informatics in Medicine (SIIM).

Dr. Kahn will present the best of research in ontologies, structured data capture, decision support, and other science focusing on data other than the images themselves. Dr. Hsu will highlight advances in radiomics, data integration and modeling, and novel healthcare applications.

Panel Organizer Statement

The panelists listed above have agreed to present the material and lead a discussion. A similar session will be presented at the annual meeting of the Radiological Society of North America (RSNA) annual meeting in December, 2016. This session is supported by the RSNA, the Society of Imaging Informatics in Medicine (SIIM), and the AMIA Biomedical Imaging Informatics Working Group.
Using Patient-Generated Data for Wellness and Biomedical Research: From Behavioral Sensing to Decision Support
Pei-Yun S. Hsueh, PhD1, Cagatay Demiralp, PhD1, Fernando Martin-Sanchez, PhD
FACMI FACMI FACHI 2, Katherine K. Kim, PhD MPH MBA3, Susan K. Peterson, PhD MPH4
1IBM T.J. Watson Research Center, Yorktown Heights, NY; 2Weill Cornell Medicine, New
York, NY, USA; 3University of California Davis, Sacramento, CA; 4University of Texas
MD Anderson Cancer Center, Houston, TX

Abstract
The rise of consumer health awareness and the recent advent of personal health IT has contributed to another shift in healthcare landscape and the accumulation of patient-generated data. Despite the rise of health consumers, the impact of patient-generated data remains to be validated. Transforming patient data for wellness and biomedical research poses many challenges. Besides the fact that patient data has been generated in large quantities that exceed what humans can consume directly, its quality and reliability vary. In this panel, we would like to review the status quo of existing evidence and present user scenarios where hybrid knowledge-augmented, data-driven methods are coupled with visualization tools to overcome barriers of interpretation and increase adaptability. We are also looking for decision support tools to better incorporate patient data for prevention and in extended care settings. This panel aims to share with leading practitioners and researchers at AMIA their efforts in using patient-generated data for wellness and biomedical research. After participating in the session, the audience should be able to assess the basic requirements for putting such data in action from the clinicians and biomedical researchers’ perspectives.

Panelists
• Prof. Fernando Martin-Sanchez, PhD, FACMI, FACHI, Weill Cornell Medicine, New York, NY.
• Prof. Katherine Kim, PhD, MPH, MBA, University of California Davis, Sacramento, CA.
• Prof. Susan Peterson, PhD, MPH, University of Texas MD Anderson Cancer Center, Houston, TX
• Cagatay Demiralp, PhD, Research Staff Member at IBM T.J. Watson Research Center, Yorktown Heights, NY.
Moderator: Pei-Yun Sabrina Hsueh, PhD, Research Staff Member at IBM T.J. Watson Research Center, NY.

Introduction
Over the last few decades, the rise of consumer health awareness has been driving the shifting of healthcare landscape, partly as a response to population aging and rise of chronic disease prevalence. In the past decade, the shift has accelerated owing to the advance of Internet and personal health management tools. More recently, a third wave of transformation is starting to take place. On the one hand, the movement of Quantified Self, e-patients, DIY healthcare, and citizen science are gaining momentum [1], with the aid from the significant improvement of health tracking capabilities on consumer devices, including mobile and wearable devices that can monitor a wide range of physiological measures (e.g., heart rate variability) and lifestyle information (e.g., physical activity and nutrition log). Polling data of 2,300 consumers shows that the adoption of mobile app for health has doubled in two years, reaching 33% in 2016, and, a majority believed that patient data is beneficial for maintaining health [2].
On the other hand, the introduction of value-based care concept has led to payment reforms. With the federal government policy initiatives such as Affordable Care Act, the focus on quality of care and the reimbursement of non face-to-face chronic care management further added fuel to the adoption of connected health technologies for self-care and just-in-time interventions. Under Medicare, 85% of provider payments are expected to tie to quality of care in 2016 [3]. It has also led to increasingly diverse channels, e.g., the nurse-led clinics and case management [4]. As a result of reconciling the two shifts, the consumer-driven healthcare is becoming a significant trend, and in turn it has driven the consumers to become pro-active. More than 90% of patients are willing to participate in shared-decision making with clinicians [5], and nearly 40% will use alternative channels for health services [2].
Despite the rise of health consumers, the impacts of patient-generated health data (PGHD) remain to be validated. Recently, some PGHD measures have been shown to correlate with the onset and progression of a variety of disease states, e.g., depression [6]. Others have shown effectiveness for patient activation in positive behavior change of physical activity and healthy eating behavior [7] and medical conditions such as obesity [8]. In the domain of disease prevention and management, researchers are moving the needle by incorporating more patient data, ranging from health and treatment history (from portal or EHR), biometric data (from lab tests and devices) to patient-reported outcomes (PRO) (e.g., survey, screening, questionnaire of medication adherence and risk assessment) [9].
In particular, the incorporation of PRO has been reported to increase patient-centered care coordination and the meaningful personalization of electronic health records. It helps assess the efficacy of treatments. For example, Geisinger Health System in northeastern Pennsylvania uses PROs to assess patients’ status before and after treatment. It can also help understand multi-domain health risk assessment (HRA) and health-related quality of life (HRQOL) [10]. In clinical trials, the use of PRO can help determine the eligibility of patients for certain clinical trials and enable comparative effectiveness on an individual basis, which becomes the basis for the N-of-1 trial framework of personalized medicine [11]. In the field of psycho-pharmacology, ecological momentary response (EMR) methods have been shown to increase the sensitivity of assessments [11] and are more amenable to the study of endpoints, as well as the determination of patient adherence and reasons for non-adherence [12]. Different frameworks are being investigated to support the generation of evidence and the subsequent design of personalized health interventions with participatory technologies (such as social media or Self-Quantification) [13,14]. However, the quality and reliability of the different PGHD data sources varies. For example, evidence is strong for the reliability of using patient health wearable data for physical activity measures, but weaker for sleep pattern and heart rate variability measures [15]. Clinical trials have been conducted to examine the clinical value of PGHD from connected health devices and PRO in different conditions, yet the results also vary widely, with some showing no improvement over standard care [16]. Further examination of the benefits of incorporating PGHD shows an increase of self-motivation to perform self-care and the timeliness of feedback for just-in-time adaptive intervention [17]. Despite that there are now initial evidence on reliability, the jury is still out there for the validity of patient-generated health data, it has been conjectured that the approach based on user-generated health data need to be coupled with incentives, incremental feedback and other strategic initiatives and careful program designs to make it work [15].

Panel Description and Learning Objectives

This panel will review the status quo of existing evidence and present success stories where hybrid knowledge-augmented, data-driven methods are coupled with visualization tools to overcome barriers of interpretation and increase adaptability. We are also looking into decision support tools in other fields to look for inspirations of possible ways to better incorporate patient data for prevention and in extended care settings. This panel aims to share with leading practitioners and researchers at AMIA their efforts to gather requirements for the next-generation healthcare delivery and population health management systems to depict the action items needed for putting patient-generate data in action. After participating in the session, the audience should be able to:

- Demonstrate how PGHD is generated through examples in participatory health and mobile health research.
- Describe emerging tools for transforming PGHD into actionable knowledge for patients and clinicians.
- Discuss gaps, drivers, barriers, and opportunities to promote the generation and application of PGHD to improve patient and clinician decision-making.

During the panel discussion, the panelists will further help address the following questions:

- How can PGHD contribute to successful provider-patient communications, risk reduction, and increase in early interventions?
- Can PGHD support shared decision making or help calibrate just-in-time intervention to patient’s values?
- Do the providers’ and patients’ beliefs and support of PGHD and approaches affect patient usage?
- Can dynamically configured healthcare IT help improve helathacer quality and patient behavior using a scalable technology-enabled platform?

Description of Speaker Topics

The panel will be organized as a series of short presentations, followed by a question and answering session in the end. The topics each speaker will cover are as follows. First, Prof. Kim will talk about patient generated data in coordination of complex chronic conditions and the emerging work in cancer and heart failure. In particular, she will report on two studies that utilized participatory and user-centered design methodologies in development of mobile platforms for care coordination for complex conditions. The first involves a personal health network for community-wide care coordination for chemotherapy, and the second focused on integrated mobile applications for heart failure. Key functionality included a shared care plan, medication management, symptom management, appointments, and integration of data and workflow with EHRs with patient-generated data incorporated into each. Considerations in designing with and for patients and implications for technology adoption and health outcomes will be discussed. Prof. Peterson will continue on reporting on the collection and use of patient-generated data in cancer prevention, clinical care, and survivorship. She will describe use cases and studies that have developed and implemented e-health and mhealth applications for behavioral assessment and intervention in several populations, including

204
individuals and families at risk for hereditary cancer, patients undergoing cancer treatment, and adolescent and young adult cancer survivors. Dr. Peterson will also describe how PGHD generated through sensor-based and mobile technology applications can guide clinical decisions and patient support for clinically relevant outcomes to cancer survivors, including therapeutic adherence, dehydration prevention, physical activity, and tobacco cessation. In addition to the role that PGHD plays in healthcare and health prevention scenarios, Dr. Martin-Sanchez will discuss PGHD can also be very useful for biomedical research purposes. Precision Medicine requires the integrated analysis of all sources of individual health data, and participatory technologies (health, social media, self-quantification) can be considered as interfaces to the Human Body and the Human Exposome (life-long exposure to environmental risk factors). Several examples of characterizing individual environmental data with mobile technologies will be presented and discussed.

Dr. Demiralp will introduce how visualization of PGHD can be applied to understand the present and future self from patient’s point of view. The goal is to help people understand and reason about their data. To that end, visualizations ease and amplify the work of cognition by re-coding information so as to exploit the perceptual abilities of the eye, which is evolved for capturing patterns and outliers. The ultimate success is to improve comprehension and decision making. In his talk, he’ll describe how data visualization principles can be easily applied to everyday visualizations. He’ll then discuss data visualization examples facilitating better decision-making by effectively conveying uncertainty and risk and present examples of using PGHD for wellness research.

**Statement of Participation**

Each of these panelists and the organizer has confirmed via email that they agree to take part in the proposed panel.

**References**

Title: Big Data for Healthcare and Life Sciences: Learning Useful Insights from Imperfect Data

Panel Chair
Jianying Hu, Ph.D.
Distinguished Research Staff Member
Program Director, Center for Computational Health
IBM T. J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598

Panelists

Nigam Shah, MBBS, Ph.D
Associate Professor of Medicine
Stanford University, 1265 Welch Road, Stanford, CA 94305

Bradley A. Malin, PhD, FACMI
Associate Professor & Vice Chair of Biomedical Informatics
Vanderbilt University, Nashville, TN USA

Patrick Ryan, Ph.D.
Senior Director, Global Epidemiology
Janssen Research and Development, Titusville NJ USA

Anil Jain, MD, FACP
Senior Vice-President & Chief Medical Officer, IBM Watson Health/Explorys
Consulting Staff, Department of Internal Medicine
Cleveland Clinic, Cleveland, OH USA

Abstract
Electronic health data has been generated and made accessible in unprecedented quantity and richness in the last decade. More recently, there has been much hype about applications of big data for healthcare and life sciences. This hype is due in part to a growing collection of success stories involving advanced data analytics bringing about real impact ranging from addressing unmet information needs for patients, clinicians, administrators and policy makers, to novel scientific research results from carefully constructed observational studies. However, at the same time, there has been much skepticism voiced citing limitations of data derived from insurance claims, electronic health records (EHRs), and other types of observational resources. The goal of this panel is to bring together experts from both academia and industry to share examples and thoughts on what precautions need to be taken in performing analytics on real world data, how careful modeling and investigation can help prevent bad science from being performed on imperfect data, and the types of useful insights that can be derived when the analysis is carried out appropriately. Learning objectives include an appreciation of both the promise and the limitations of observational health data, and practical guidance on how to leverage advanced analytics to harness useful insights from such data.

General Description
Traditional medical research approaches are hypothesis-driven and based on deductive reasoning carried out on relatively small amounts of data collected in a highly controlled manner, such as
randomized clinical trials. With the arrival of big data there is increasing recognition that many types of research questions and information needs can be addressed through inductive learning that starts with observations and develops models that lead to novel insights. Such approaches open up vast new possibilities of discovery and also help circumvent the confirmation bias that limits hypothesis-driven investigations. The key, however, is to understand what types of questions can be best addressed with observational studies and how to test the consistency of the results and ensure the validity of the conclusions.

This panel will seek to address a range of related questions including:

- What is useful and how does it depend on the end user? For example, for research, useful means advancing our knowledge about the basis of medicine, whereas in operations, useful means making sure the right patients get the care they need at the right time.
- How should we measure data quality and potential distortion from point of care to final analytical results?
- What data standardization is preferred to address heterogeneity in populations and the data capture process for the purpose of phenotyping and analytics?
- How do we facilitate and measure reproducibility across databases or research networks?
- How can we validate and expand known associations into new discoveries?
- Can we leverage longitudinal modeling to smooth out irregularities in data?
- Under what conditions can subphenotyping strategies uncover mixtures of groups and why imperfect is not always incorrect?

Prof. Shah will examine the use of aggregate patient data in an informatics consult in a research setting. A few years ago, researchers at Stanford demonstrated how data from a research clinical data warehouse could be used to drive evidence-based medicine (http://www.nejm.org/doi/full/10.1056/NEJMp1108726). Prof. Shah will discuss recent progress towards helping clinicians leverage aggregate patient data for decision-making at the point of care. He will illustrate the idea with a demo of a search engine built to support such an informatics consult, and review how decisions about data pre-processing affect phenotyping as well as downstream analyses that estimate treatment effects from imperfect data.

Prof. Malin will present research conducted at Vanderbilt University on discovery of complex phenotypes, stratification of patients, and predictive modeling using data from multiple large academic healthcare systems including Northwestern Memorial Hospital and Vanderbilt University Medical Center. He will illustrate how they were able to discover inconsistencies in vocabulary usage and subsequently design techniques to define core phenotypic topics that enabled an alignment of disparate patient populations. He will also review how longitudinal modeling, in combination with risk stratification, can enhance stability in the prediction of blood pressure control. The presentation will conclude with a discussion on how data privacy protections could be considered a form of generating imperfect data and how appropriate design of such mechanisms can enable both validation and discovery.

Dr. Ryan will discuss the evolving role of observational data analysis in generating and disseminating real-world evidence about disease, health service utilization, and the effects of medical products. He will highlight the role of the pharmaceutical industry in establishing and promoting best practices in the appropriate use of observational data by conducting methodological research, developing standardized analytics tools and demonstrating successful applications to important clinical questions.
He will illustrate how three core methodological issues – reproducibility, generalizability, and heterogeneity- represent the primary challenges to observational health data sciences, and how collaborating in open source communities with other stakeholder groups offer a viable path forward.

Dr. Jain will discuss the opportunities and challenges of utilizing electronic health record data and claims data generated from healthcare systems that are undergoing payment and delivery transformation. Using examples from implementations at several health systems, the presentation will show how real-world data exhaust generated by these systems can both accelerate net new knowledge discovery and support population health focused data sciences to drive innovations as academic and non-academic health centers are faced with diminishing margins. Despite advances in supporting interoperability between information systems, the ongoing need to further curate real-world data will be discussed. The presentation will conclude with a demo visualizing how this imperfect data can be used to predict health care utilization in the setting of population health management.

**Explanation of Urgency of Topic**

The promise of big data for healthcare and life sciences lies not in the generation of the data itself, but in how it is used. We are marching towards bigger, broader, and more complex systems with more fine-grained detail on patients coming from diverse data sources. At the same time, we are increasing the distribution and reuse of data beyond its initial primary setting and context of collection. As data science becomes more popular, there will be substantial interest in applying arrays of unsupervised and supervised machine learning models to this data. However, the extent to which the results are reliable and useful will be contingent on recognizing bias, inaccuracy in data collection (and subsequent filtering), and using multiple modes of evidence to obtain consistency in discoveries and predictions. There is thus an urgent need in the community to have an informed and thoughtful discussion around the methodological considerations needed to realize the potential of big, heterogeneous, imperfect data.

**List of Questions**

1. What should be the role of observational evidence compared to other sources of evidence?
2. How much confidence can be taken from observational analysis?
3. Will initiatives in reproducible research and transparency resolve concerns in such investigations?
4. How do you validate findings from imperfect data?
5. How do you access data quality, and how do you factor that into your analysis?
6. How do you evaluate utility of the insights (AUC vs. Net reclassification for example).
7. How can we instill trust in the findings discovered from imperfect big health databases?
8. Under what circumstances can discoveries made in one setting be translated into another?

**Statement of Participation:**

All participants have agreed to take part on the panel.
Standardizing Research Common Data Elements: Initiatives, Exchange Formats and Current Use by Patient-Level Trial Results Databases

Vojtech Huser¹, Murat Sincan², Dikla Shmueli-Blumberg³, Rachel Hess⁴
¹Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD
²National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD
³EMMES Corporation, Rockville, MD ⁴University of Utah, Salt Lake City, UT

Abstract

Sharing of de-identified patient level data from clinical trials is increasingly becoming a norm and this trend is increasing the importance of research Common Data Elements (CDEs). CDEs have the potential to increase the value of trial data by making it easier to integrate data across multiple trials. CDEs can reduce study start-up costs, improve the quality of collected data, and facilitate cross study comparisons, data aggregation and meta-analyses. This panel will (1) describe current CDE initiatives, highlight several trial data sharing platforms (including their use of CDEs), and describe CDE informatics standards; (2) describe a case study a site using REDCap Electronic Data Capture system and their CDE experience; (3) describe a case study of a trial data sharing platform of the National Institute on Drug Abuse; and (4) showcase a multi-site observational study that first selected suitable CDEs and implemented collection of trial data directly within an Epic EHR.

General Description (including ‘Why the topic of this panel is timely’)

Sharing of de-identified patient level data from clinical trial is increasingly becoming a norm to ensure reproducibility of analyses, increase public trust in research enterprise, and to avoid wasting research resources. To further facilitate comparison and aggregation of data from multiple research studies, use of Common Data Elements (CDE) is increasingly encouraged by public research funders, such as the National Institutes of Health. CDEs have the potential to increase the value of data collected in clinical trials by making it easier to integrate data across multiple trials. CDEs can reduce study start-up costs, improve the quality of collected data, and facilitate cross study comparisons, data aggregation and meta-analyses. Several trial data platforms now have multiple year experience with hosting and distributing trial data, including the effort to use medical terminologies and CDEs for semantic data integration.

This panel will describe the current state of the clinical research informatics of CDEs and showcase examples of their use in clinical research studies and trial results databases.

Speakers:

Vojtech Huser will introduce the topic of common data elements. He will describe a clinical trial lifecycle and impact of CDEs during the trial design stage and impact of CDEs on secondary use of data (during data sharing). He will provide an overview of data platforms that allow individual patient level data from clinical trials and the current adoption of CDEs by trial data platforms. See Table 1 for overview of trial data sharing platforms ordered by size (number of trials with data). In the second part, Dr. Huser will give a brief overview of several CDE initiatives (e.g., NIH initiatives: PhenX, PROMIS, cancer CDEs; non-NIH initiatives: CDISC Therapeutic Area (TA) standards or FDA [Food and Drug Administration] standards, namely Study Data Tabulation Model and Define-XML standard). Finally, he will also describe where CDEs can currently be downloaded and what informatics standards are being used to import and export empty Case Report Forms (CRF) or form-collected patient data (e.g., National Library of Medicine (NLM) CDE repository at https://cde.nlm.nih.gov, REDCap shared library at https://redcap.vanderbilt.edu/consortium/library or University of Muenster’s repository at http://medical-data-models.org).

Murat Sincan will describe an implementation of a REDCap Electronic Data Capture system at the National Institute of Dental and Craniofacial Research (NIDCR) and how principal investigators at NIDCR are currently impacted by CDEs. He will provide an EDC user perspective on informatics issues with importing CRFs with common data elements from the NLM’s CDE Repository or other external repositories. He will describe how CDEs can be incorporated into project CRFs and the practical issues that arise during the lifecycle of a project with an emphasis on change management and data harmonization.
Dikla Shmueli-Blumberg will approach the topic of CDEs from the perspective of a trial results database. She will describe the National Institute on Drug Abuse (NIDA) Data Share website. The website provides a platform for timely sharing of de-identified research data from NIDA Clinical Trials Network (CTN) studies after their completion. In addition to the data files (available as ASCII or SAS format), descriptive metadata, research protocols, annotated case report forms, and listings of assessments used in the studies are also provided. This effective and user-friendly method of dissemination of results to the larger research community has been effective at promoting secondary data analyses and providing valuable information to the field of substance use treatment.

Rachel Hess will describe efforts of the PaTH PCORNet Clinical Data Research Network (CDRN) to integrate patient-reported outcomes into research embedded within clinical care. The team worked with clinical, patient, and operational stakeholders to identify and integrate PROMIS and disease-specific quality of life measures. The PaTH team chose PROMIS instruments because of their validity across a wide variety of disease states and future ability to use computer adaptive testing to make questionnaires even briefer. PaTH clinical stakeholders identified disease-specific instruments that had high clinical relevance, would contribute to advancing the understanding of the experience of that condition, and were practical within routine care. Patient stakeholders validated the domains of interest and ensured that measures chosen were reflective of their experience of their health condition. Operational champions worked closely with research and technical teams to deploy questionnaires in Epic’s MyChart personal health record. This allowed information to flow into the electronic health record and be used for clinical care and research. The PaTH experience served as a proof-of-concept for two health systems (participating in PaTH) to integrate patient-reported outcomes into clinical care. These teams used local resources as well as Epic’s Boost program to ensure timely implementation. For individuals without personal health record access (either because they chose not to sign up or their institution did not have questionnaire capability in the personal health record), questionnaires were available via REDCap or paper. All paper questionnaires were entered into REDCap by staff; REDCap-based responses were not available for clinical care.

Panel type: Didactic; All participants have agreed to take part on the panel
Anticipated audience: clinical research informatics researchers, clinical study PIs, research clinicians, EDC users and administrators, research coordinators, medical terminology experts, data scientists reusing data

Discussion points:
1. What is the current adoption of Common Data Elements and does it differ by domain (e.g., cancer, asthma, stroke, rare diseases)?
2. How can different CDE initiatives be harmonized (avoid creating multiple CDEs for similar research concepts)?
3. Are existing CRI standards sufficient to import a form (or a set of CDEs) from one system to another? What is the current adoption of these standards by different electronic data capture systems?
4. What is the relationship of CDEs (that try to conceptualize clinical research) to biomedical terminologies (that try to conceptualize routine medical care [e.g., SNOMED])?
5. Is it currently possible to track actual usage of CDEs across medical domains? How long after introduction of CDEs can we demonstrate that CDEs are truly adding some value to the research enterprise?

List of speakers: (1) Vojtech Huser, Staff Scientist at National Library of Medicine; member of the CDE Working Group of the Trans-NIH BioMedical Informatics Coordinating Committee (BMIC); (2) Murat Sincan, Staff Scientist at National Institute of Dental and Craniofacial Research (NIDCR), provides informatics leadership to the NIDCR REDCap server within NIDCR intramural research program; (3) Dikla Shmueli-Blumberg, Project Manager at EMMES Corporation for NIDA Clinical Trials Network; (4) Rachel Hess, Professor of Population Health Science and Internal Medicine, University of Utah School of Medicine, representative of the PaTH PCORnet Clinical Data Research Network

References

Table 1: Overview of data sharing platforms for clinical trials’ patient-level data

<table>
<thead>
<tr>
<th>Initiative</th>
<th># of Trials*</th>
<th>Launch Year</th>
<th>Website</th>
<th>Link to an Example Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PHARMA ORIGINATED</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Sphere</td>
<td>54</td>
<td>2014</td>
<td><a href="https://www.projectdatasphere.org">https://www.projectdatasphere.org</a></td>
<td><a href="https://www.projectdatasphere.org/projectdatasphere/html/content/137">https://www.projectdatasphere.org/projectdatasphere/html/content/137</a></td>
</tr>
<tr>
<td><strong>NIH SPONSORED</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ACADEMIC/INDUSTRY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OTHER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Extending Commercial Electronic Health Record Systems through Interoperable Applications and Services: Experiences and Lessons Learned from Four Leading-Edge Institutional Programs

Kensaku Kawamoto, MD, PhD, MHS¹, Scott P. Narus, PhD, MS¹,², Richard A. Bloomfield Jr., MD³, Alistair Erskine, MD⁴, Blackford Middleton, MD, MPH, MSc⁵

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT,
²Intermountain Healthcare, Murray, UT,
³Duke University Health System, Duke University, Durham, NC,
⁴Geisinger Health System, Danville, PA, ⁵Apervita, Inc., Chicago, IL

Abstract

Commercial electronic health record (EHR) systems such as Epic and Cerner are increasingly offering powerful interoperability frameworks for extending their capabilities through interfaced applications and software services. These interoperability frameworks include application programming interfaces (APIs) for reading and writing data from the EHR, embedding Web applications seamlessly into the native EHR user interface, and consulting external clinical decision support Web services from within the EHR’s native rules engine. Importantly, several of these APIs utilize standards-based, vendor-agnostic interfaces such as the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) data interface. In this panel, leaders from four leading healthcare organizations will describe how their institutions have established operational programs for developing applications and services that extend the capabilities of their EHR systems to improve clinical care. Example applications and services that have been developed by the panelists’ organizations include an integrated rheumatology management application; patient-facing medication compliance applications; a pediatric growth chart application; and a sepsis predictive modeling service. The panelists will also describe some of the key challenges and lessons learned, solicit additional insights from the collective experience of the audience, and provide guidance for colleagues seeking to extend their EHR systems at their respective organizations.

Description

The panel will be organized as follows:

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>Middleton</td>
<td>Opportunity and landscape introduction</td>
</tr>
<tr>
<td>15 min</td>
<td>Kawamoto</td>
<td>University of Utah experience</td>
</tr>
<tr>
<td>15 min</td>
<td>Narus</td>
<td>Intermountain Healthcare and Healthcare Services Platform Consortium experience</td>
</tr>
<tr>
<td>15 min</td>
<td>Bloomfield</td>
<td>Duke University Health System experience</td>
</tr>
<tr>
<td>15 min</td>
<td>Erskine</td>
<td>Geisinger Health System experience</td>
</tr>
<tr>
<td>25 min</td>
<td>All</td>
<td>Panel discussion with audience</td>
</tr>
</tbody>
</table>

Dr. Middleton will serve as the moderator and introduce each of the panel members and their organizations. Dr. Middleton will provide an overview of the exciting opportunity that has been made available by many commercial EHR vendors to extend their capabilities through interfaced applications and services. This overview will be followed by presentations by the panelists, who will describe why their organizations have made a commitment to interoperable applications and services; their strategy, governance, and approach; their progress to date and roadmap ahead; and lessons learned. Each panelist will also showcase novel capabilities developed by their organizations through these programs. These presentations will be followed by a panel discussion with the audience moderated by Dr. Middleton.
Bloomfield: Dr. Bloomfield is Director, Mobile Technology Strategy and Assistant Professor of Internal Medicine and Pediatrics at Duke Health. Starting in the Spring of 2014 Dr. Bloomfield and the team at Duke began exploring the integration of the SMART and FHIR APIs with their Epic EHR. It was determined that this was not only feasible, but could help accelerate development and innovation throughout the health system. By January 2015 Duke had built the required infrastructure to make this a reality, and was able to demonstrate several SMART/FHIR apps (open source and proprietary) running within the Epic desktop environment as well as within the provider-facing mobile apps. This infrastructure based on FHIR Draft Standard for Trial Use Release 1 (DSTU1) was approved for production use in August 2014, becoming the first use of FHIR in a production Epic system. Since that time, the infrastructure has been updated to DSTU2 and work is ongoing to add support for patient-facing use-cases as well as write support. This work will be complete in March 2016 and is part of Duke’s commitment to the White House from the summer of 2015 with respect to patient-facing use cases for precision medicine. Dr. Bloomfield is currently working on ways to streamline the process for app integration to enable this technology to quickly scale. He will present an overview of this work, lessons learned, and give a demo of the technology.

Erskine: Dr. Erskine is Chief Clinical Informatics Officer at the Geisinger Health System. He is responsible for sequencing future and innovative technologies at Geisinger and for harmonizing data across the clinical care, research and health plan enterprise. Dr. Erskine heads the Division of Applied Research and Clinical Informatics (DARCI) which engages staff in the design and configuration of Geisinger’s clinical information systems and vendors for extension and associated standards.

Kawamoto: Dr. Kawamoto is Associate Chief Medical Information Officer, Assistant Professor of Biomedical Informatics, and Director of Knowledge Management and Mobilization at the University of Utah. He also co-leads the University of Utah’s Interoperable Apps and Services (IAPPS) initiative, which is a multi-stakeholder effort to enable standards-based, interoperable applications and software services to improve health and health care. Dr. Kawamoto will describe why the University of Utah initiated the IAPPS program, its governance model and resourcing, and the inter-disciplinary collaboration of groups including the Epic EHR team, biomedical informatics, the computer gaming graduate program, and the Center for Medical Innovation. He will discuss the initiative’s central strategy of enabling various groups from within and outside the organization to innovate and contribute, as well as the tactics being pursued to enable such crowd-sourced development while ensuring security and maintainability. He will also describe OpenCDS (www.opencds.org), an open-source clinical decision support framework being leveraged within this initiative, and the initial set of provider and consuming-facing applications and services that have been developed, such as for “gamifying” medication compliance for patients and integrating a sepsis prediction module into the native EHR rules engine. Dr. Kawamoto will also discuss key lessons learned, such as the tension between EHR-agnostic interoperability and the desire to leverage EHR-specific, more advanced APIs; and the need for additional semantics to be defined within FHIR to enable true interoperability.

Narus: Dr. Narus is a Medical Informatics Director and Chief Clinical Systems Architect at Intermountain Healthcare. He also leads and participates in technical direction for the Health Services Platform Consortium (HSPC). Dr. Narus will discuss Intermountain’s efforts to partner with Cerner to develop an open, standards-based service layer on Cerner’s EHR infrastructure in order to allow broader capabilities for in-house and third party application innovation. He will describe internal efforts to develop and implement applications on this new service layer, and lessons learned. He will also describe Intermountain’s participation in HSPC, which is seeking to expand the idea of open, standards-based services and application interoperability across EHR and healthcare organization boundaries.

Middleton: Dr. Middleton is Chief Informatics & Innovation Officer at Apervita, Inc., the provider of the largest marketplace for the creation, exchange, and use of data and analytics in healthcare. He is immediate Past-Chair of the Board of Directors of the American Medical Informatics Association (AMIA), and Instructor in the Harvard TH Chan School of Public Health in the Departments of Health Policy and Management, and Policy Translation and Leadership Development. Previously, he was a Professor of Biomedical Informatics, and/or of Medicine, at Stanford, Harvard, and Vanderbilt Universities, and he held executive leadership roles at MedicaLogic/Medscape, Partners Healthcare System, and at Vanderbilt. Dr. Middleton’s work is focused on clinical informatics – the applied science surrounding strategy, design, implementation, and evaluation of clinical information systems in complex environments. Dr. Middleton will provide an overview of the opportunity for innovating and extending the functionality of EHR systems. He will also provide a brief overview of the main mechanisms offered by EHR vendors for extension and associated standards.
evolves Geisinger’s facilities to take advantage of ultramodern technologies. Dr. Erskine will describe how his team is using SMART and FHIR to contribute to the emerging healthcare Web App ecosystem in an EHR-agnostic manner. Dr. Erskine will describe the unique functionality developed at Geisinger, such as a workflow-optimized and visually rich rheumatology Web application (https://gallery.smarthealthit.org/xg-health-solutions/enrg-rheum). In addition, he will describe the technical approach used, including getting FHIR from Geisinger’s interface engine (Rhapsody v6.2), from middleware (Sansoro Emissary product), from Epic Corp (through its available native FHIR support), and from Cerner’s Healthe Intent platform. Dr. Erskine will also describe how Geisinger has organized a FHIR-based dev/ops process branded “SynrG”, which is designed to incrementally invest in home-grown applications based on merit/outcomes and compliance with standard FHIR-based data resources. Dr. Erskine will also discuss the many lessons learned through the Geisinger journey in interoperable applications and services.

**Middleton:** Dr. Middleton will lead a moderated discussion with the audience. The objectives of this discussion will be to answer questions from the audience, engage in a stimulating exchange of ideas, and gain further insights on the topic from audience members’ collective experiences. The discussions questions listed below will be used to stimulate this discussion.

**Significance of panel topic and anticipated audience**
Whereas home-grown EHR systems served as the platforms for innovation in biomedical informatics for many years, there is a clear trend for home-grown systems to be replaced by commercial EHR systems. Thus, it is imperative that biomedical informaticists become experts of commercial EHR systems and how to innovate within and in conjunction with those systems. This panel will help audience members working through this transition to better understand how biomedical informaticists can thrive and make essential contributions by leveraging EHR systems’ currently available and continually improving frameworks for interoperability.

**Discussion questions**
What experiences have you had with extending your EHR system’s functionality, and what lessons have you learned?
What recommendations do you have for standards development and implementation in this area?
What gaps do you see in the current interoperability frameworks offered by EHR systems?

**Participation statement**
All proposed panelists have agreed to participate in the panel if the proposal is accepted.
Report from the ED-WG Frontline: Biomedical and Health Informatics
Baccalaureate (BHIB) Course

Saif Khairat, PhD, MS1, Glynda Doyle, RN, MSN2,
Indra Neil Sarkar, PhD, MLIS3, Jeff Williamson, MEd4,
1University of North Carolina, Chapel Hill, NC; 2British Columbia Institute of Technology,
Vancouver, Canada; 3Brown University, Providence, RI; 4AMIA, Washington, DC,

Participants
• Organizer and Panelist: Saif Khairat, PhD, Carolina Health Informatics, University of North Carolina
• Panelist: Glynda Doyle, RN, MSN, British Columbia Institute of Technology
• Panelist: Neil Sarkar, PhD, MLIS, Center for Biomedical Informatics, Brown University
• Panelist, Moderator: Jeff Williamson, MEd, AMIA

Abstract
Biomedical Informatics training at the graduate and post-graduate level has expanded to include almost all clinical specialties. Whether it is a Masters, PhD, or the fellowships in the Sub-Specialty in Clinical Informatics, graduate students or returning professionals have various informatics training options. However, informatics education at the undergraduate level is not as well established. Despite initiatives such as TIGER1 and the Health Information Technology Scholars Program2, as well as a limited number of undergraduate degree programs (e.g., at Arizona State University3), there is still a need to provide interprofessional informatics education at the baccalaureate level.

The Education WG has endorsed and supported undergraduate education as its priority since AMIA 2014 panel4. Education WG members have been leading efforts to develop a Biomedical and Health Informatics course that addresses fundamental concepts and core competency skills. Panelists aim to further report on the expectations of the AMIA community with regards to using an undergraduate, interprofessional informatics course. The goal is to present students with an overview of the role of biomedical informatics in clinical, research, and operation practices. This panel will announce the completion of AMIA’s first undergraduate course geared towards various health and IT specialties including medicine, nursing, and allied professions.

Learning Objectives
• Discuss the scope and content of the BHIB course
• Reflect on core undergraduate informatics competencies (general and specialty specific)
• Learn about on-going and future efforts to integrate informatics in baccalaureate education
• Discuss feedback on usability of the course nationally and internationall
• Identify and invite qualified Informatics professionals to keep content updated as a sustainability plan
• Provide a report summarizing course development since AMIA 2014

Audience
Due to its timeliness and relevance, this session will be of interest to all health and biomedical informatics educators.

Background and report of discussions and feedback from AMIA 2014 Symposium interactive panel
Biomedical Informatics training has expanded to include almost all clinical specialties. Whether it is a Masters, PhD, or fellowship in Sub-Specialty in Clinical Informatics, graduate students or returning professionals have various informatics training options. However, biomedical informatics education at the baccalaureate level is not as well established (with very few notable exceptions, like Arizona State University3). Despite initiatives such as TIGER, ONC HIT Workforce Development, the Health Information Technology Scholars Program, and others, there is still a need to provide interprofessional education at the baccalaureate level. As the field of biomedical informatics becomes increasingly relevant in myriad healthcare settings, aiming to leverage more and complex and data, there is
a growing need to introduce fundamental concepts to undergraduate students to enhance their knowledge base and stimulate interest in career paths in biomedical informatics.

The AMIA Education Working Group (ED-WG) has been leading efforts to develop an Interprofessional Informatics course for Baccalaureate healthcare students that address fundamental informatics concepts and core competency skills. The ED-WG has built an undergraduate course that includes core informatics competencies for various clinical specialties including medicine, nursing, allied health professions, and computer science. The proposed course is designed to serve as a foundation for faculty, students, and other educators striving to increase biomedical informatics literacy and basic knowledge in a realm where much confusion exists about the domain. This effort provides fundamental knowledge of biomedical informatics to pre-med, nursing and allied health majors, as well as computer and information science undergraduate students. The course is designed to serve as a foundation for health professionals, which is different from the advanced ONC-funded Health IT Workforce Curriculum Components. Moreover, this course would complement the current AMIA efforts towards high school and post-graduate education, and promote the development of a “pipeline” of biomedical informatics professionals.

**Explanation of proposed course/modules and proposed process**

The AMIA ED-WG developed a course that includes 18 modules that cover fundamental and core biomedical informatics competencies. We invited AMIA professionals to develop these modules as per their area of interest. Each module will be offered as voice-over PowerPoint’s with learning objectives, content, questions and answers, and case studies relevant to all healthcare professions. The AMIA 10x10 courses provide a good example and possible framework for this program.

The modules could either be taken as a full course (15-16 weeks), or individual modules could be pulled from the program and be used by an educator in their individual course. For example, a nursing instructor teaching an Ethics course could integrate the “Ethical/Legal Issues in Health Informatics” into their course. Or, a Pharmacology instructor could use the “Precision Medicine” module in their course. The program would be applicable and available internationally (including countries that biomedical informatics may have a practical impact, e.g., Malaysia, India, and South Africa). It is a goal to make this program available as an Open Education Resource.

**Proposed Modules:**

1. What is Biomedical Informatics?
2. Tools of Biomedical Informatics
3. Terminologies/vocabularies/data sets
4. Computation
5. Clinical application
6. Ethical/legal issues
7. Research methods/ evidence-based practice
8. Telehealth
9. Global health informatics
10. Public Health Informatics
11. Mobile health technologies
12. Consumer health informatics
13. Human Factors
14. Personalized Medicine
15. Digital Literacy
16. Precision Medicine
17. The future of Biomedical Informatics
18. Careers in Biomedical Informatics

**Engage AMIA audience in discussion about:**

- Suggestions around course usability and circulation
- Potential sustainability plans
- Possible dissemination channels

**Panelist Presentations**

Saif Khairat, PhD, has been involved with Health Informatics since 2006. He is assistant professor at the Carolina Health Informatics Program and School of Nursing at the University of North Carolina- Chapel Hill. Dr. Khairat’s research focuses on the intersection of EHR usability and patient safety. Dr. Khairat was co-Principal Investigator to a federal grant award from the Health Resources and Services Administration (HRSA) entitled “Telehealth Resource Center Grant Program”, at the University of Minnesota. Dr. Khairat earned his PhD in Health Informatics at the Informatics Institute at the University of Missouri with a focus on ICU clinical communication. During his Informatics training, Dr. Khairat worked as a Research Fellow at the Division of Clinical Informatics at Harvard.
Medical School. Dr. Khairat is lead author to numerous publications and serves as a scientific reviewer to national and international conferences and journals.

Dr. Khairat is currently the Chair of the Education Working Group at AMIA, member of the AMIA Working Group Steering Committee, and in the past, he has served on Student Working Group committees. Dr. Khairat will talk about the course structure, the need for highly trained informaticians to participate in content development, and the need for content reviewers. He will pose questions to the audience related to the need for an undergraduate informatics course.

Glynda Doyle, RN, MSN teaches at the British Columbia Institute of Technology (BCIT) in Vancouver, British Columbia. She completed her MSN at the University of British Columbia in 2011. Ms. Doyle is focused on the integration of informatics into the BCIT Bachelor of Science in Nursing (BSN) and Specialty Nursing curricula. She is particularly interested in the role of mobile technologies and their impact on nursing student’s clinical judgment and decision making.

Ms. Doyle is a co-investigator in several inter-disciplinary research projects within BCIT and also in collaboration with other Canadian nursing schools studying the impact of mobile devices laden with clinical resources, social networks and e-portfolios on nursing students and their education. Ms. Doyle recently co-developed an open source education resource/textbook (Clinical Procedures for Safer Patient Care) funded by BC Campus. Glynda is a peer leader in BC for the Canadian Association Schools of Nursing and Canada Health Infoway Digital Health Faculty Peer Leader Network. Ms. Doyle will discuss the national and international need for a course such as this at the undergraduate level and its value to the field of informatics and the AMIA community.

Neil Sarkar, PhD, MLIS, is director of the Center for Biomedical Informatics at Brown University with broad interests in biomedical informatics research and education. He is also the immediate past chair of the AMIA Education Committee, where he first identified the need for AMIA to develop a comprehensive pipeline for a biomedical informatics workforce, akin to how other STEM-C disciplines have approached similar challenges. His research focuses on the development of methodologies for linking biomedical data across heterogeneous sources, leveraging work in ontologies, information retrieval, and natural language processing, with particular application to the study of complex disease phenotypes (e.g., pre-term birth) as well as identification of putative new medical knowledge (e.g., medicinal plants). He has developed full semester courses as well as short (one-week) immersion educational experiences that cover fundamental introductory and methodological concepts biomedical informatics, which have been taken by both undergraduate, graduate, and post-graduate learners. Dr. Sarkar will describe his experiences in teaching traditional courses in biomedical informatics that includes undergraduate students, and offer a broad perspective of how such courses are essential in the development of a pipeline of biomedical informatics professionals.

Jeffrey J. Williamson, M.Ed., Vice President of Education and Academic Affairs, will moderate this session. He has been with AMIA since 1998. Jeff’s portfolio of programs and activities at AMIA’s includes oversight of annual conferences, e-learning activities, the American College of Medical Informatics, the Academic Forum, and providing continuing education benefits for members. He is the lead staff member responsible for Clinical Informatics Board Review Course, certification and accreditation. Mr. Williamson will discuss AMIA’s vision and endorsement for undergraduate education, logistics around course availability and hosting, and AMIA’s sustainability plan for the BHIB course.

All panelists agree to participation if the panel is accepted.

References

2. HITS: Health Information Technology Scholars Program. Available at: http://www.hits-colab.org

217
The Medication-use Process: Current Challenges and Potential Solutions

Wing Liu, PharmD; Lori Idemoto, MS, RPh; John Poikonen, PharmD; Sarah Alameddine, PharmD; Scott D. Nelson, PharmD, MS

1Vanderbilt University Medical Center, Nashville, TN; 2Swedish Medical Center, Seattle, WA; 3Avhana Health, Cambridge, MA; 4Nova Southeastern University College of Pharmacy, Palm Beach Gardens, FL

Abstract

The medication-use process (MUP) is complex, often misunderstood, and can have serious downstream consequences in the accuracy and communication of medications, resulting in patient safety concerns or harm. It is imperative that informaticists understand the MUP to minimize potential adverse events and unintended consequences. The panel will highlight a patient’s journey, focusing on the areas where their healthcare delivery is impacted by the nuances of the MUP. The audience will experience how interoperable medication lists, medication order timing, electronic controlled substance prescribing, and their socio-technical aspects provide opportunities for health information technology to improve patient safety and process efficiency.

Learning objectives:
1. State some of the common obstacles to sending and receiving accurate medications lists.
2. Explain strategies for improving and interpreting interoperable medication lists.
3. Identify variables in their institution that may cause timing discrepancies.
4. Formulate an approach to develop solutions that prevent timing discrepancies.
5. List e-prescribing enhancements that can improve provider and community pharmacy communication.
6. Describe obstacles to electronic prescribing communication enhancements.
7. Learn and describe strategies that can be used to reduce the impact of medication-use process errors.
8. Gain a greater awareness of the socio-technical issues that impact the electronic medication-use process.

General Description

The medication-use process is a complex integration of prescribing, perfecting medication orders, preparing, dispensing, administering, monitoring, and educating patients about medications. There is an unmet need for informaticists to not just simply understand the context surrounding the medication-use process, but also participate in continuously improving its nuanced challenges. Due to the complexity of the process and growing regulatory requirements related to medication management, health information technology can play a significant role in improving patient safety and process efficiency. Therefore, it is imperative that informaticists understand the medication-use process and engage in a scientific approach to minimize potential adverse events and unintended consequences. The objective of this panel is to describe a patient journey and the various stages of the medication-use process they touch, with each presenter putting a spotlight on a particular aspect of its subsequent safety or efficacy ramifications. Challenges and considerations of the medication-use process will be highlighted, such as socio-technical aspects, medication order timing, electronic controlled substance prescribing considerations, and interoperable medication lists. The presenters for this panel are pharmacists, informaticists and experts in the medication-use process that will share some of their experiences, insights, and potential solutions related to these medication-use process challenges.

Importance

Understanding some of the current and common challenges of the complex medication-use process is critical for informaticists in order to improve patient safety and care. Participants will get value from learning how to organize and access health information data using available technology and processes, and understand how to apply it to improve practice patterns. The intended audience will be informaticists and clinicians that are unaware of the medication-use process, and experienced informaticists that would like to apply the strategies described to their own institutions. Participants will be encouraged to share their experiences and views, and will have opportunity for group discussion questions, provide feedback, and question the panel. Focusing on the patient journey and effectively leveraging technology in the medication-use process provides the essential context for delivering higher quality and safer care for all patients.
Moderator and introduction (Scott Nelson)
The medication-use process will be introduced along with a brief introduction to the patient journey.

Medication lists, searching for the truth (Sarah Alameddine)
Over the past few years, Meaningful Use requirements have established terminology standards for communication between EHR systems to promote health information exchange. This exchange of documents has enabled providers to share important patient information, such as medications the patient may be taking. However, the day-to-day use of these medication lists has raised many concerns. Interoperable medication lists are often an inaccurate reflection of the source, as they commonly contain free text medications, absent or inaccurate medication end dates, imprecise RxNorm codes, and unreliably documented “patient reported” medications. Understanding the gaps and inconsistencies in the medications lists is an important part of using this important medication information particularly in the monitoring and educating medication use process, effectively. This presentation will discuss some of the common challenges of and some strategies for improving the usability of interoperable medication lists.

Timing is everything (Wing Liu)
As one of the medication administration "five rights", medication timing is critical for patient safety; however, actual medication administrations typically do not occur as ordered in CPOE. While CPOE can reduce some errors, it may have difficulty with the functionality and logic surrounding medication timing and cannot fully predict the clinical intent of the user. These limitations often result in timing discrepancies, such as ‘now/stat’ orders being administered too close to the next scheduled dose, system limitations to hold medication doses, or limitations in addressing the temporal relationship of a patient’s temporary health status (such as pregnancy). This presentation will focus on the application of informatics in the design, development, and implementation of decision support and guideline systems pertaining to the various challenges of medication timing in the administration phase of the medication-use process.

Controlled substances and e-prescribing (John Poikonen)
Electronic prescribing (e-prescribing) has increased significantly across the nation, and this electronic communication between prescribers and pharmacies may reduce medication errors and increase patient safety. However, not all changes to medication therapy are communicated between prescribers and pharmacies, such as prescription cancelations form the prescriber or notification of prescription pickup from the pharmacy. Additionally, there are many issues that must be taken into consideration when e-prescribing and dispensing controlled substances. Implementation of Prescription Change transactions, Cancel Prescription transactions, electronic prescribing of controlled substances, and prescription drug monitoring programs could support improved communication and feedback with the medication-use process in the ambulatory setting. Although standards for these electronic prescribing enhancements have been established, implementation may be delayed based on existing functionality of EHRs and community pharmacy dispensing systems. This presentation will discuss various challenges and obstacles to the electronic prescribing process and these prescribing communication enhancements and the pharmacists role in “perfecting” and dispensing in the medication use process.

Socio-technical considerations (Lori Idemoto)
The rapid adoption of electronic health records (EHRs) and computerized provider order entry (CPOE) have changed the roles and ways healthcare professionals care for their patients and interact with each other. While these informatics tools have made improvements to the medication-use process, they have also introduced unintended consequences and new types of errors that have negatively impacted patient safety. Changes in workflow and communication along with an over dependence on technology have created false expectations and results in the emergence of new errors and an increase in the frequency of specific medication use process errors. This presentation will focus on socio-technical changes that may adversely impact the medication use process.

Discussion Questions
1. In your experience, how accurate do you think the medication lists are that you receive?
2. What are some of the challenges you or your organization have experienced in medication list accuracy?
3. How do you or your organization use external medication lists (CCDA/MU) for medication reconciliation?
4. What medication timing issues have you or your organizations encountered in your clinical systems?
5. What approaches have you or your organizations used to solve the medication timing issues mentioned?
6. If your organization allows medications to be placed on hold, what does the process look like and what issues have you encountered?
7. Has your organization experienced a patient safety event caused by a patient continuing to take a medication that was intended to be discontinued?
8. How does your organization communicate discontinued medications to primary care providers or pharmacies?
9. If your organization electronically prescribes controlled substances, what has your experience been implementing such a feature?
10. What approaches have you or your organizations taken to mitigate the impact of these socio-technical issues related to the medication-use process?
11. What socio-technical issues have you or your organizations observed affecting the medication-use process? How are they similar or different to those described here?

**Participation statement**
The authors listed in this panel have agreed to take part in the panel.

**References**
Clinical Informatics Fellowship – Future Workforce Development

J Edward Maddela, MD¹, Umar Iqbal, MD², Yumi DiAngi, MD³,
Erika Green, MD⁴, Daria Ferro, MD⁵

¹,²University of Arizona College of Medicine, Phoenix, Arizona; ³Stanford University and Hospital Systems, Palo Alto, California; ⁴Regenstrief Institute Center For Biomedical Informatics, Indianapolis, Indiana; ⁵The Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania

Abstract

Clinical Informatics (CI) is the newest addition to the American Board of Medical Specialties (ABMS), with its first board certification issued in 2013 to 456 physicians¹. The American Medical Informatics Association (AMIA) was first to recognize the importance of formal physician training to meet the growing need for a CI workforce. AMIA has designed both the Core Content and Program Requirements for the new subspecialty training programs in 2008². Today, 20 Accredited Council for Graduate Medical Education (ACGME) fellowship programs exist nationally with more programs expected to join. Little is known about the structure of the various training programs. In this didactic panel we will: 1) review examples of four ACGME accredited CI fellowship programs, 2) seek opinions from healthcare organization and academic medical center leadership, and other key stakeholders in physicians' CI education concerning the skillset needed for graduating fellows, and 3) discuss the challenges of the programs as they evolve to educate and train the next generation of CI physician workforce.

General Description of Panel

As a new and unique subspecialty, Clinical Informatics (CI) arose from the need of physicians to be integrally involved with the appropriate use of clinical information systems. Beginning in 2018, physicians who seek Board Certification in Clinical Informatics must complete a 2-year Accredited Council for Graduate Medical Education (ACGME) Clinical Informatics Fellowship. Designed to prepare fellows for a career in CI, these programs provide them with the skills necessary to support informatics enabled solutions of clinical services in various settings. With this didactic panel, we will present and discuss the diverse curricula among four ACGME Accredited CI Fellowship programs: The Children’s Hospital of Philadelphia, Regenstrief Institute, Stanford Children’s Hospital, and the University of Arizona – College of Medicine in Phoenix. After a brief introduction, each program will spend 10 minutes presenting their distinctive curriculum and 9 minutes seeking feedback and/or addressing questions from the audience. The final 10 minutes of this panel presentation will be open for a general discussion.

Intended Audience

The intended audience of this panel includes but is not limited to Clinical Informatics fellows, hospital and academic medical center leadership, CI Fellowship Program Directors, CI Fellowship Faculty, physicians seeking CI Board Certification, residents, medical students, non-physician clinicians, and the general audience interested in the education of the future CI physician workforce.

Explanation of Topic Urgency

This topic is urgent and timely because physicians pursuing board certification in CI need to decide if an ACGME CI fellowship or the Practice Pathway is more suitable for them. Even though information about these fellowships can be accessed through their websites, specific details about and comparisons between each fellowship program are not readily available. Additionally, a formal discussion about the structure and content of these fellowships has yet to take place on a national platform. With many more fellowship programs seeking ACGME accreditation, we anticipate the dialogue that will arise from this didactic panel presentation will bring value to the intended audience.

Brief Description of Presentations

The Children’s Hospital of Philadelphia (CHOP)

The CI Fellowship Program at CHOP is now in its second year, led by ten Faculty Physician Board-certified Clinical Informaticists. The program is housed in the Department of Biomedical and Health Informatics (DBHi), CHOP’s home for the development of innovative solutions to healthcare's immediate and long-term informatics needs. CI
Fellows at CHOP are exposed to and receive support from not only Clinical Informaticists, but also those working in a variety of areas of Biomedical Informatics, leading to interdisciplinary and enriching experiences.

CHOP’s CI Fellowship curriculum is composed of core clinical rotations as well as longitudinal leadership and research fundamentals based primarily at CHOP. Didactic training takes place through master’s degree level courses, online (Oregon Health & Science University) and locally (the University of Pennsylvania), in addition to regular small group sessions (e.g. journal club). There is purposeful flexibility designed in the schedule to allow exploration and professional development for each fellow based on individual objectives and career goals. This professional development is supported by close personalized mentoring, and is considered by the fellows to be one of the highlights of the program at CHOP.

In this presentation, we will explore the general structure of CHOP’s CI Fellowship curriculum and highlight the importance of an individualized approach to each fellow’s career goals in order to maximize the benefits of CI Fellowship training.

**Regenstrief Institute, Center for Biomedical Informatics**

Regenstrief Institute has established its Clinical Informatics fellowship program prior to ACGME accreditation and is currently lead by JT Finnell MD MSc FACEP. Regenstrief Institute has a rich legacy in informatics research dating back to the late 1960’s. It has developed one of the nation's first electronic medical record (EMR) system and computerized provider order entry (CPOE). It is well known for development of the nation’s first coding standard for laboratory tests and clinical measurements LOINC® and the Indiana Health Information Exchange.

The presentation will include an overview of the current structure of the program, research opportunities with Regenstrief faculty, several options to work with our partners that vary from public health institutions to private tech companies, and opportunities for fellows to gain operational experience in two health systems using Epic and Cerner.

**Stanford Health Care, Stanford Children’s Hospital, and Stanford University**

This is the first ACGME accredited clinical informatics program nationally. It has now accepted its third year of fellows. Leadership and support for the program comes from both the adult and children’s hospital, the University, Santa Clara Valley Hospital System (safety net hospital system), and industry.

The presentation will:

1) Cover core rotations, elective, clinical, and leadership opportunities at the University and health system
2) Highlight operational and research activities of the fellows
3) Communicate challenges and evolving program changes
4) Discuss post fellowship job opportunities

**University of Arizona, College of Medicine-Phoenix**

This two-year fellowship in CI is comprised of eight 12 week quarters spread across 5 major teaching institution hospitals in Phoenix, Arizona. The first 2 quarters are divided into 6 months of core basic knowledge – similar to that of basic sciences in medical school: Architectural Overview, Safety & Quality, Process Improvement, Clinical Decision Support, Privacy & Security, and Business Processes. Quarters 3 and 4 involve a 3-month short term project, followed by a 1-month elective, and then a 2-month data analytics rotation. Quarter 5 starts with a 2-month quality improvement project. The remainder of the time is spent in community and self chosen electives. Quarter 6 to 8 is heavily focused on the final long term project, which entails an initiation, execution, and reporting phase respectively. All projects have a written deliverable and presentation. Every Wednesday, fellows spend the day at the University of Arizona campus where there are monthly activities including didactic sessions presented by faculty, fellow led journal club, grand rounds, and monthly reviews. Concurrently, fellows are obtaining core content knowledge via Oregon Health & Science University’s Clinical Informatics Fellowship Education Programs which consists of eight courses totaling 24 credits resulting in a Graduate Certificate in Biomedical Informatics. Fellows also spend 15 percent of their time providing direct patient care in their specialty area in order to maintain clinical skills and deepen their experience in clinical environments.
Learning Objectives

After participating in this session, the learner should be better able to:

<table>
<thead>
<tr>
<th></th>
<th>Bolster the general structure of CI fellowship. This includes the core rotations, content of curriculum, research opportunities, clinical time, elective opportunities, national involvement, and leadership opportunities as well as resources needed to support the fellow.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Define the acquired skill set and expected competencies of a CI fellow. This includes listing job opportunities following fellowship and the skill sets that would prepare the fellow for those positions.</td>
</tr>
<tr>
<td></td>
<td>Describe possible roles that the CI fellow can support within a health system. This includes leadership on informatics tasks, sharing lessons learned from health system projects, and supporting change management and innovation within health systems.</td>
</tr>
</tbody>
</table>

Discussion Questions

1. What were your interests in Clinical Informatics prior to starting your fellowship?
   a. How has the fellowship supported, shaped, shifted these interests thus far?
2. How do you envision yourself using your CI training in your professional career?
3. What opportunities has the fellowship afforded that you would not likely experience learning CI as a practitioner without formal training?
4. In what ways can CI fellows act as ambassadors of the field?
5. For those already working in CI leadership roles, what are the skills that you think are the most needed for a Clinical Informatician?
   a. Do you perceive that the fellowships are training individuals to meet these needs?
6. Please provide the top 3 learning objectives that a CMIO would see as critical in graduating CI fellows?
7. What level of importance and training should CI fellowships place in the subject areas of business processes, organizational finance, governance, budgeting, and strategy?
8. Which kind of roles do you foresee CI fellows assuming after graduation?
   a. How many years of experience would anticipate before seeing them in leadership positions if they started fellowship straight from residency?
9. Do you anticipate or encourage other disciplines starting similar formalized training programs – i.e. nursing, pharmacists, physician assistants, and how could these disciplines collaborate or take lessons learned from current CI fellowship programs?
10. What kind of involvement should fellows training in ACGME accredited CI fellowship programs have with AMIA?
11. What do you know now that you wish you knew before you started the CI fellowship?
12. What advice would you give to someone considering a fellowship in CI?

Statement from Panel Organizer

All presenters listed above agree to participate in this Didactic Panel. - J Edward Maddela, MD

References

Didactic Panel: Interactive Systems in Healthcare

Lena Mamykina, Ph.D1, Maddhu Reddy, Ph.D2, Katie Siek, Ph.D3, Gabriela Marcu, Ph.D4, Leslie Liu, Ph.D3.

1Columbia University, New York, NY, USA; 2Northwestern University, Chicago, IL, USA, 3Indiana University Bloomington, Bloomington, IN, USA, 4Drexel University, Philadelphia, PA, USA

Abstract

In this didactic panel, we will review an emerging and growing field of investigation within medical informatics, interactive systems in healthcare. A set of panelists with complementary areas of expertise and substantial experience in biomedical informatics and human-computer interaction will provide an overview of the field, highlight examples of cutting-edge investigations, discuss grand challenges, and outline training and career opportunities. The information presented during the panel will be informed by the Workshop on Interactive Systems in Healthcare (WISH 2016) co-hosted with the SIGCHI conference on Human-Factors in Computing Systems, CHI 2016. Given the growing interest in this area, we expect the panel to be educational and allow for a lively discussion of topics relevant to AMIA members.

Introduction

Health Information Technology (HIT) has the potential to transform healthcare by positively impacting quality, efficiency, and cost-effectiveness. Similarly, new self-monitoring technologies and data science methods open unprecedented opportunities to inform decisions and choices of individuals who must manage their health and chronic diseases at home. However, despite ongoing efforts by many government agencies, HIT continues to experience low levels of adoption\(^1\). Moreover, a growing body of research questions its impact on medical care\(^2\), for example by highlighting the unintended consequences of HIT\(^3\), and medical errors that result from poorly designed computing systems\(^4\). Researchers have argued that many of these negative consequences result from a mismatch between the reality of conducting clinical work and the structure of computing applications that are meant to support it\(^5\). Similar concerns are raised in regards to the usefulness and longevity of technologies for health and wellness\(^6\).

To address these limitations and remove barriers to the successful adoption of HIT, new research initiatives are focusing on a better alignment of HIT and clinical practices, and an approach to system design that is informed by best practices in Human Factors and Human-Computer Interaction (HCI). However, these efforts currently exist in several disjointed research communities, without established pathways for transfer of knowledge and expertise. These communities include but are not limited to Biomedical Informatics, HCI, Computer Science, Social Sciences, and Behavioral Medicine, among others. Each of these fields has its own venues for disseminating research results that rarely overlap. Therefore, researchers and practitioners interested in designing patient and clinician-centric HIT have little opportunity to interact and develop a shared body of knowledge across these communities. As a consequence, there exists a potential to create deeper and more profound connections among the biomedical, informatics, human-computer interaction, medical sociology and anthropology communities that would lead to the development of new methods, approaches, and techniques for removing barriers to the adoption of HIT.

In response to these challenges, the American Medical Informatics Association (AMIA) and the Association of Computing and Machinery (ACM) have co-sponsored a series of Workshops on Interactive Systems in Healthcare (WISH). The overarching goal of these workshops is to establish lasting and meaningful connections between these communities, and bring together students and researchers from various disciplines who are working on creating, implementing, and evaluating innovative health technologies. The first workshop in the series, WISH 2010 was co-hosted with CHI 2010 in Atlanta, GA. The workshop was sponsored by the NSF, ACM, SIGCHI, and Microsoft Research. It attracted close to 150 participants and included both invited panels and peer-reviewed technical program. From 2011-2014, WISH was co-hosted with the Annual Symposium of the American Medical Informatics Association. In 2016, WISH is co-hosted with CHI as a two-day workshop, including an open symposium during the first day, and an invitation-only meeting of the WISH steering committee during the second day.

In this panel, we will outline a broad research agenda for Interactive Systems in Healthcare and provide highlights from WISH 2016.
Panelists and their Qualifications

The panelists are well-known researchers in the fields of Human-Computer Interaction, Computer Science, and Biomedical Informatics who bring their expertise and interdisciplinary perspective to advance the agenda of the workshop. Three of the organizers have served as co-chairs for previous WISH (Drs. Mamykina and Reddy in 2011, Dr. Siek in 2012) that were co-hosted with AMIA. All co-organizers are prominent members of HCI and Biomedical Informatics communities and have served as members of the program committees for CHI and AMIA for multiple years.

*Lena Mamykina, PhD,* is an Assistant Professor of Biomedical Informatics in the Department of Biomedical Informatics at Columbia University. Her primary research interests reside in the areas of Biomedical Informatics, Human-Computer Interaction, Ubiquitous and Pervasive Computing, and Computer-Supported Collaborative Work. Dr. Mamykina’s broad research interests include individual and collective cognition, sensemaking and problem-solving in the context of health, health care, and health management.

*Dr. Madhu Reddy, PhD,* is a Professor of Health Communication in the Department of Communication Studies and a core faculty member in the Center for Communication and Health at Northwestern University. Previously, he was in the College of Information Sciences and Technology at Penn State University and helped start Penn State’s Center for Integrated Healthcare Delivery Systems. Dr. Reddy’s primary research interests are at the intersections of Medical Informatics and Computer Supported Cooperative Work (CSCW).

*Katie Siek, PhD,* is an Associate Professor in Informatics at Indiana University Bloomington. Her primary research interests are in human computer interaction, health informatics, and ubiquitous computing. More specifically, she is interested in how sociotechnical interventions affect personal health and well being. She has actively participated in the CHI community (PC member 2012 and 2013) and AMIA community (PC member 2011 and Vice Chair for the AMIA Annual Symposium 2013) and other conferences and workshops integrating computing and health.

*Gabriela Marcu, PhD,* is an Assistant Professor of Human-Centered Computing in the College of Computing and Informatics, and a Research Fellow with the A.J. Drexel Autism Institute, at Drexel University. She holds a Ph.D. in Human-Computer Interaction from Carnegie Mellon University, and a B.S. in Informatics from the University of California, Irvine. Dr. Marcu’s research seeks to improve coordination and collaborative reflection in health and social services through technology.

*Leslie S. Liu, PhD,* is a postdoctoral fellow in Informatics at Indiana University Bloomington. Her primary research interests lie within the intersection of human computer interaction, computer supported cooperative work, and health informatics. She both studies and designs technologies for patient populations.

In the proposed panel, Dr. Liu will serve as a moderator, and Drs. Mamykina, Reddy, Siek, and Marcu will serve as panelists.

Proposed Topics

We propose the following general topic for discussion during the panel:

1. **Overview of the field and main areas of investigation (Dr. Mamykina)**

In the first presentation, Dr. Mamykina will review a history of interactive systems in healthcare and how transformations in interactive technologies impacted healthcare practice. In addition, Dr. Mamykina will review recent work on advancing interactive systems in healthcare and outline prominent areas of investigation. These include, but are not limited to, mobile applications for health and wellness self-management, interactive systems for clinicians’ communication, interactive information displays for patient care teams, among many others.

2. **Representative presentations from WISH 2016 illustrating contemporary work within the field (Dr. Marcu)**

To illustrate contemporary research in interactive systems in Healthcare, Dr. Marcu will provide an overview of the five original manuscripts presented during WISH 2016. These manuscripts will be selected based on their scientific merit and also to represent the diversity of investigations.

3. **Grand challenges (Dr. Siek)**

Dr. Siek will highlight grand challenges for interactive systems in healthcare identified during the meeting of the WISH 2016 steering committee. These challenges will outline opportunities for new research as well suggest
priorities for funding agencies. In addition, Dr. Siek will discuss the vision for future Workshops on Interactive Systems in Healthcare developed during WISH 2016.

4. Training and career opportunities (Dr. Reddy)

In the last presentation, Dr. Reddy will discuss different skills necessary to conduct impactful research in interactive systems in healthcare and opportunities for fostering these skills within academic institutions and industry-oriented training programs.

After the initial presentations, we plan to open the discussion to the audience. We will begin by asking the panelist to comment on the following set of general questions:

1. What unique opportunities and challenges exist in relation to design methods as well as evaluation methods appropriate for health-oriented interactive systems?
2. What opportunities for innovation in interactive systems in healthcare exist within the current landscape of Health IT?
3. What opportunities for funding and publishing work related to interactive systems in healthcare exist within the different research communities?

In addition, we expect new questions from the audience that will explore a variety of topics related to interactive systems in healthcare.

Conclusion

Given the growing interest to interactive systems in healthcare and growing adoption rates of Electronic Health Record systems, we believe this panel will provide a valuable overview of this area to both newcomers to AMIA as well as to members who are conducting active research and develop new interactive tools for improving health and delivery of healthcare.

All panelists agreed to take part in the panel.

References


doi:10.1145/2750858.2807554
Building a Research Ecosystem upon the Early Success of ResearchKit

Panel Moderator
Kenneth D. Mandl, MD, MPH
Boston Children’s Hospital
Harvard Medical School
Boston, Massachusetts

Panelists
Richard Bloomfield, Jr., MD
Duke University
Durham, North Carolina

Deborah Estrin, PhD
Cornell Tech
New York, New York

Stephen Friend, MD, PhD
Sage Bionetworks
Seattle, Washington, Washington

Pascal Pfiffner, MD, PhD
Boston Children’s Hospital
Harvard Medical School
Boston, Massachusetts
Abstract
In March of 2015, Apple introduced ResearchKit, an apps framework and tool chain for conducting clinical studies that begin when subjects discover and interact with a smart phone app. ResearchKit enables investigators to readily inform and consent patients into medical and health studies. Once a subject is enrolled, ResearchKit facilitates collection of personal health data, device data and patient reported outcomes (PRO) from distributed populations. The framework launched with five initial studies that immediately demonstrated unprecedented brisk enrollment of subjects “in the wild.” Because ResearchKit is open source, extensions and adaptations began immediately. This panel explores new use cases and platforms for the technology. Topics for presentation and discussion include the role of ResearchKit in Precision Medicine, the value of connecting HealthKit to the research enterprise, transposition of the framework to the Android platform, connection of ResearchKit to the i2b2 ecosystem, and alternative enrollment pathways, including “prescribing” ResearchKit apps to selected subjects in a trial.

General Description of Program
Accrual to clinical trials has traditionally been an expensive, time consuming, and labor-intensive process. The process has been fraught, in that consents are often difficult to understand, electronic data capture from patients often uses customized and unfamiliar solutions, researchers lose contact with subjects, and biometric measurement on patients is time consuming and costly.

Because hundreds of millions of people carry smartphones in their pockets, an alternate paradigm becomes possible. If a patient can be made aware of a study, the pathway to enrollment is as easy as opening up a smartphone app. ResearchKit commoditizes the introduction to research studies, instantiates a crystal clear consent process, facilitates both active and passive collection of device data, and makes reporting of PROs trivially straightforward.

The initial release of ResearchKit has been focused on subjects “in the wild” finding a study and choosing to participate. But direct-enrollment of subjects is only one of several study designs that could be supported by the platform. Further, while ResearchKit arose on the iOS platform, an analog for Android -- ResearchStack extends the reach of the approach to much larger populations.

Statement of Need
This panel will be held about 18 months after the initial release of ResearchKit. We suspect that a significant representation within AMIA will be using it creatively. This panel brings together key leaders and innovators advancing the capabilities and possibilities for smartphone-based trial designs who will bring participants up to the cutting edge, and also catalyze a discussion and perhaps even identify new directions for the project within our community.

Kenneth D. Mandl, Director of the Boston Children’s Hospital Computational Health Informatics Program, and a Professor of Biomedical Informatics at Harvard Medical School, will begin with a brief overview of the initial work done in ResearchKit and paint a picture of how it could be widely extended within the biomedical and health research enterprise and specifically for precision medicine. Each panelist will then be invited to make a presentation in an area of his or her expertise for 10 minutes. The total presentation time will be kept to under one hour to allow lively discussion with the audience.

Stephen Friend, President of Sage Bionetworks, and a creator of ResearchKit, will motivate the creation of ResearchKit and through the lens of the mPower study of Parkinson’s disease, discuss the power of app-based enrollment in clinical trials, collection of digital phenotypes using smartphones, and open data.

Ricky Bloomfield, Assistant Professor of Pediatrics, Duke University, will discuss experience with a pediatric ResearchKit study Autism and Beyond; a ResearchKit study of children between 1 and 6 years demonstrating the feasibility of at-home video facial analysis for screening of childhood mental health disorders. While a prior, Duke-only study had recruitment of about 100 individuals over a 12-month period, after release of the ResearchKit app there were over 250 individuals consenting to the study in just the first weekend alone, with almost 1200 in less than 5 months.

Deborah Estrin, Professor of Computer Science at Cornell Tech and a Professor of Public Health at Weill Cornell Medical College, will discuss the transposition of ResearchKit to ResearchStack, an SDK and user interface framework for building research study apps on Android. ResearchStack will help developers and
researchers who have existing apps on iOS more easily adapt those apps for Android. The Research and ResearchStack SDKs have similar, though non-identical functionality, but the a common framework and naming scheme are hoped to greatly speed up adaptation of ResearchKit apps to Android and vice versa.

Pascal Pfiffner, Faculty in the Boston Children’s Hospital Computational Health Informatics Program and Lecturer at Harvard Medical School, will talk about the marriage of ResearchKit and Fast Healthcare Interoperability Resources (FHIR) in the C3-PRO tool chain, and how this combination simplifies survey delivery and response collection in a standardized, portable fashion. Additionally, the C3-PRO tool chain not only handles end-to-end research data delivery from participants’ smartphones to i2b2, it also enables researchers to “prescribe” a research app to patients already in the health system, extending (research) data collection to the mobile domain.

Questions for audience discussion

● Have any of you designed ResearchKit apps?
● Have you participated in a ResearchKit study?
● Have you extended the ResearchKit open source framework
● What digital phenotypes might you define with smart phone data collection?
● What types of studies would you like to see enabled?

Statement from the Organizer

Each participant has agreed to take part on the panel.
Secure Record Linkage for Precision Medicine and Patient Centered Outcomes Research

Daniella Meeker, PhD1, Toan Ong2, Xiaoqian Jiang, PhD3, Abel N. Kho, MD, MS4, Jason N. Doctor1

1University of Southern California, Los Angeles, CA; 2University of Colorado, Anschutz Medical Campus, Aurora, CO; 3University of California, San Diego, La Jolla, CA; 4Northwestern University, Chicago, IL

Abstract

The ability to link records across multiple sources enables richer data for a wide variety of applications in patient-centered outcomes research and precision medicine. The standard implementation of a linkage system uses personally identifying information from different sources to perform a probabilistic or deterministic identity match, with manual inspection to resolve uncertainty. While there are established commercial record matching systems for treatment purposes and Health Information Exchange, regulations, risks and risk tolerance differ in the context of research. Furthermore, if existing systems can be improved upon from the standpoint of both performance and privacy protection, disclosures required for manual adjudication can be limited. Each of the panelists will describe their experiences with implementing a privacy-preserving record linkage (PPRL) approach, and then discuss the pragmatic implications, and share results of a comparative evaluation. The methods discussed will include using seeded one-way hashing, Garbled Circuits, multi-party GMW (Goldreich-Micali-Wigderson), Intel® Software Guard Extensions, weighted and order-preserving encryption methods, and approaches using count queries across parties to infer a total ordered data from partial orders. We will describe how different data elements, weighting strategies, and order-preserving hash functions can impact match performance. Panelists will directly address issues associated with the HIPAA Privacy Rule and implementation of different methods in the presence or absence of a trusted broker. The panel will present considerations for privacy, security, trust, accuracy, and computing resources that merit attention, giving attendees an overview of the different feasible options for privacy-preserving record linkage.

Introduction

Patient Centered Outcomes Research and Precision Medicine share the challenge that a single entity rarely holds all of the information necessary to conduct meaningful research. Creating a collection of patient-centered records that includes data from multiple entities distributed over space and time that can be used for research is a significant challenge. Frequently, consent is raised as a solution to privacy risks that limits liability and compels record release. While a virtuous first step, consent does not eliminate the need to link identities across independently managed data systems in a reliable and secure way. The identity of the individual providing consent still needs to be resolved and validated before records are released – a person volunteering for research participation recruited from the internet must be linked to a definitive identifier when she requests that her health records are released for research. Furthermore, consent reduces liability of data holders, but consent does not protect privacy.

Technical barriers to protect participant and institutional privacy are coupled with concerns of legal liability and privacy risks. In absence of universal identifiers, record linkage relies on matching methods that typically require use of personally identifying information. Effective research must meet competing requirements for precision, privacy, security, efficiency, and data completeness. While there are widely adopted commercial solutions in use by service providers of all types, research presents distinctive challenges. Research is in certain settings an optional activity - data holders have no obligation (and often minimal incentives) to release records for research. For outcomes research and precision medicine, records are typically requested in large batches rather than one-by-one, as it is the case in health information exchange. The extent to which health information exchange solutions easily translate into research solutions requires weighing of different desired features.
Implementers of record linkage systems have a variety of options for architecture, algorithms, procedures, and data management. Matches can be made in deterministic or probabilistic ways – selecting decision thresholds for probabilistic methods may favor sensitivity or specificity. Which data elements are employed, how they are weighted in matching algorithms, and how they are cleaned and processed to account for misspelling, keystroke errors, and aliases all impact both the performance of the match and the likelihood that privacy may be compromised. Encryption methods that do not preserve string lexicography or numeric order may provide simple solutions but limit flexibility for identifying matches. Also, implementers must consider whether it is a greater risk to create a persistent identifier that is reused over time or if it is desirable to recreate fresh linking identifiers with every data use. Discrepancies in organizational policies, computing or data quality resources may impact how different parties coordinate their efforts and assess feasibility. Finally, purpose for use must be considered – the cost of mistakes in matches for an observational data analysis is typically substantially lower than a mistake that results in patient contact or disclosure. Panelists will describe how the methods in question perform not only on metrics related to matching, but on the broader range of pragmatic issues necessary for implementation.

Best practices for data sharing in research should minimize avoidable disclosures. These practices include using analytic methods that remove requirements for centralizing data, replacing identifying information in research records with encrypted identifiers, etc. Performing privacy-preserving record linkage is a frequently neglected topic that is crucial to enable joint analytics. We will address this critical issue in this panel.

Panelists

Abel N. Kho, MD will present a pragmatic system in use by multiple collaborating health systems in the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN). He will describe the procedures for data cleaning, preprocessing, and hashing of patient identifiers, which removes all protected health information prior to matching. He will describe the rationale for selecting data elements used for matching – including combinations of dates of birth, names, social security numbers, and soundex variants. The method employed an intentionally conservative deterministic approach – minimizing false positives. Dr. Kho will describe the common organizational concerns raised and steps taken to enable real-world implementation.

Toan Ong, PhD will describe novel algorithms and procedures that address challenges associated with missing data in probabilistic clear-text and privacy preserving record linkage. In contrast to more conservative deterministic methods, Dr. Ong will present algorithms designed to maintain computational efficiency while maximizing the accuracy of weighted probabilistic linkage mechanisms. The first approach Weight Redistribution removes fields with missing data from the set of quasi-identifiers and redistributes the weight from the missing attribute based on relative proportions across the remaining available linkage fields. Distance Imputation imputes the distance between the missing data fields rather than imputing the missing data value. Linkage Expansion adds previously considered non-linkage fields to the linkage field set to compensate for the missing information in a linkage field. The rationale for the methods and constraints applied in the algorithms will be described. Comparison of performance of these methods to the deterministic approaches for both clear text and encrypted data will be included.

Xiaoqian Jiang, PhD will present an approach that eliminates requirements for trusted brokers and compare security and computational efficiency. His presentation will compare three secure multi-party computation (MPC) algorithms for record linkage. These approaches do not require centralized architectures and the privacy of each party can be guaranteed. The Garbled Circuit (GC) based patient linkage algorithm and will be compared to a method based on Intel® Software Guard Extensions (SGX) and another method based on Goldreich, Micali, and Wigderson algorithm. The relative flexibility and computational efficiency of these approaches will be described.

Jason N. Doctor, PhD will describe methods that do not require sharing row-level data with a broker and rely on (1) shared encryption algorithms and (2) infrastructure that many research networks have implemented that allow for querying the number of patients meeting a specific criteria and sharing of count statistics with restrictions on the size of cells that can be shared. He will describe how sorting algorithms are related to minimizing the number of queries required to assign and identify overlapping members between sites. He will describe the number of data elements that are required to support this approach and the privacy risks that might be introduced in a probing attack.
Discussion Questions

1. These methods have been implemented in heterogeneous networks including the VA, FQHCs, and Academic Medical Centers – how do the policies of different institutions/environments influence design and implementation?

2. How can these methods facilitate Open Data and Open Science initiatives for PCOR and Precision Medicine?

3. Should these methods be applied in a way that allows them to persist over time across a research network, or should new linkages be computed for each data use? What are the privacy and efficiency considerations?

4. How much of security is derived from the different encryption methods vs. limiting access control?

5. What is the best way to communicate to a regulatory expert the pros and cons of these different methods?

6. How should privacy concerns associated with these approaches be presented to an IRB or a research participant weighing the risks and benefits of participating?

Statement of Participation

All panelists have agreed to participate in the panel.

References


Panel Proposal

Calibration of Predictive Models for Clinical Decision Making:

Personalizing Prevention, Treatment, and Disease Progression

Lucila Ohno-Machado, MD, PhD,1 George Hripcsak, MD, MS,2 Michael Matheny, MD, MPH,3 Yuan Wu, PhD4, Xiaoqian Jiang, PhD1
1Dept. Biomedical Informatics, University of California San Diego, La Jolla, CA
2Dept. Biomedical Informatics, Columbia University, New York, NY
3Dept. Biomedical Informatics, Vanderbilt University, Nashville, TN
4Dept. Biostatistics, Duke University, Durham, NC

Abstract

A large number of clinical decision support applications rely on predictive models for binary outcomes. These models usually estimate the probability for an outcome of interest (e.g., probability of breast cancer). Given that clinical decisions rely on these probability estimates, it is critical that they be well calibrated (i.e., the estimates are sufficiently close to the true underlying probability that the event will occur). Surprisingly, the evaluation of predictive models rarely includes an assessment of calibration [1–3], while essentially all include an assessment of discrimination (e.g., AUC: areas under the ROC curve). In this panel, we will describe the perils of using non-calibrated models for making clinical and administrative decisions and describe various methods to recalibrate existing models or to include calibration in model development algorithms. We will conclude with a discussion of future directions for research on model calibration.

General Description of the Panel and the issues

Predictive models are used in clinical practices to generate individualized estimates for medically relevant outcomes. Examples include NCI’s model for breast cancer [4], MELD for end-stage liver disease [5], APACHE score for ICU mortality [6], the Framingham calculator for coronary heart disease risk [7]. These models often deal with binary outcomes (e.g., event=1 means death and event=0 indicates survival). Output probability estimates for the event of interest are typically produced. Differently from typical classification algorithms that focus on the discrimination (i.e., whether cases receive higher estimates than controls), the estimated probabilities of outcomes matter in a clinical context because these estimates are often used to determine individual risk for a particular patient, and can guide important clinical decisions like the initialization of anti-lipid pharmacotherapy for an individual at high risk for cardiovascular disease [8] or referral for chemoprevention trials or radical mastectomy for a woman with high chances of developing breast cancer [9]. Therefore, we need to measure the calibration of the estimate or prediction on an individual level, by checking how close the prediction is to the underlying probability for each patient. Because we cannot evaluate this for a single patient (i.e., a stratum with \( n=1 \)), certain proxies are necessary. For example, we can assess the probability of the event in a group of “similar” patients to determine whether individualized estimates are well calibrated or not. Adequate calibration is often coupled with adequate discrimination but a highly discriminative classifier (e.g., a classifier with a large AUC [10]) is not necessarily well calibrated. The consequences for an individual can be quite impactful hence it is very important for predictive models to be well calibrated. The panel will focus on various aspects of calibration for predictive models in the healthcare context.

Structure of the Proposed Panel

The goal of this panel is to provide an overview of recent research in the field of predictive model calibration. The panel will include four recognized researchers in biomedical predictive modeling. Each of these researchers will address a different aspect of calibration in predictive modeling and describe their experience with different applications.
The panel will be hosted by Dr. Lucila Ohno-Machado, Professor of Medicine and Chair of the Health System Department of Biomedical Informatics at UCSD. For many years she has been working on assessing the quality of individual estimates in decision support systems [11–13]. She will provide an overview of the opportunities and challenges in developing calibration techniques for clinical decision support.

The first panelist will be Dr. George Hripcsak from Columbia University. Dr. Hripcsak is Chair and Vivian Beaumont Allen Professor of Biomedical Informatics. He has extensive experience with predictive models and temporal analysis [14–16]. He will share his insights on related calibration issues.

The second panelist will be Dr. Michael Matheny. He is Assistant Professor of Biomedical Informatics and Biostatistics from Vanderbilt University, and studied discrimination and calibration for clinical applications [17,18]. Dr. Matheny will share his insights of discrimination and calibration in post-procedural mortality in interventional cardiology studies.

The third panelist will be Dr. Xiaoqian Jiang. He is an Assistant Professor in the Health System Department of Biomedical Informatics at UCSD. Dr. Jiang has developed algorithms for calibration of predictive models [19–21] and is currently employing them in various contexts. He will provide a comparison of different calibration techniques and contextualize their application.

The fourth panelist will be Dr. Yuan Wu from Duke University. Dr. Wu is Assistant Professor of Biostatistics and Bioinformatics. He has been working on survival analysis and calibration of predictive models using statistical tools [22]. He will share his experience from a statistician’s perspective.

**Urgency of the Topic and Relevance for AMIA**

Data-driven predictive models serve as important tools for informaticians, biostatistician, epidemiologists, and clinical researchers. They are used for personalized prevention, treatment and clinical outcome prediction [23–25]. This panel addresses the evaluation of important building blocks for clinical decision support: predictive models for binary outcomes such as hospital readmission and mortality estimates. A large portion of AMIA members are involved in these activities as users, developers, or educators. The anticipated audiences include, but not limited to healthcare policy makers, biomedical researchers, as well as other stakeholders (e.g., patients, clinicians, etc.)

**Discussion questions**

Potential discussion questions include, but not limited to:

- How to measure calibration?
- What are consequences of non-calibrated models?
- How can we improve state-of-the-art risk prediction models (risk scores in intensive care units, renal transplantation, congestive heart failure) using calibration?
- Can we build models that exhibit high discrimination but poor calibration or vice-versa?
- How to recalibrate models so they can be applied in a particular population?
- What are the pros and cons of different calibration methods?
- How to build models with calibration guarantees?

**Statement from the Organizer**

All of the panelists have approved the writing of this proposal and have agreed to take part on this panel.

**References**


Electronic case reporting: 360° perspective by public health, informatics, and healthcare stakeholders

Catherine Staes, BSN, MPH, PhD\textsuperscript{1}, Shu McGarvey, CBAP, CSM\textsuperscript{2}, Shan He, PhD\textsuperscript{3}, Ryan Arnold, MPH\textsuperscript{4}, Laura Conn, MPH\textsuperscript{5}

\textsuperscript{1}Biomedical Informatics, Salt Lake City, Utah; \textsuperscript{2}Northrup Grumman, Atlanta, GA, \textsuperscript{3}Intermountain Healthcare, Salt Lake City, UT, \textsuperscript{4}Houston Health Department, Houston, TX, \textsuperscript{5}Centers for Disease Control and Prevention, Atlanta GA

Abstract

This panel brings together key perspectives on the challenges and strategies used for public health case reporting, a Meaningful Use (MU) requirements for the electronic health record incentive program recently included in MU Stage 3. As public health and health care enterprises, and electronic health record vendors prepare for electronic case reporting, there are questions about the proposed strategies, current functionality, and readiness of the systems required to move from vision to reality. This panel includes presentations that will allow the audience to understand the need and systems from the public health (receiving) perspective, the informatics methods and tools being leveraged, and early experiences with performing case reporting from a health care enterprise. Questions will be posed for 30 minutes of discussion at the end to garner the perspectives and input from an audience of informatics experts working with the systems and technologies required to realize electronic case reporting.

Background

Electronic Case Reporting (eCR) is a priority activity for moving public health surveillance from a predominantly manual and paper-based process to an electronic one. In fact, the Center for Medicare and Medicaid has expanded the Meaningful Use requirements for the electronic health record incentive program to include case reporting in Stage 3.\textsuperscript{1} eCR is complex, and requires changes to process, policy, and technology for both healthcare and public health systems. The complexity and segmented funding for eCR requires a divide and conquer approach with work distributed across multiple public health partners. This presentation brings together the perspectives of three major stakeholders, and presents a consolidated picture that illustrates the eCR vision.

The topic of this panel is timely, particularly for the target audience attending AMIA. Many healthcare institutions grapple with the unfunded, yet legally required, mandate to report information to public health authorities when patients present with particular conditions. It is well-documented that reporting of communicable diseases is incomplete and untimely, and providers often are not aware of reporting requirements.\textsuperscript{2} Advances in technology and information standards (both semantic and syntactic) enable eCR, but are not sufficient. Several threads of enabling work are currently underway, and advances in clinical decision support and knowledge management make the vision for eCR possible. The widespread use of electronic health records in both inpatient and outpatient care\textsuperscript{3}, further increase the opportunities for eCR. Public health reporting can move from a largely manual, fax- and phone-based process to an ongoing, systematic and automated process that manages most of the reporting requirements and supports information exchange during public health emergencies. Systems are needed to manage communication of common conditions under surveillance (e.g., vaccine-preventable diseases, such as measles and pertussis) as well as rare, emerging or re-emerging events (e.g., ebola or zika virus).

Currently, several federal agencies and local partners are collaborating to develop infrastructure and knowledge resources to implement eCR during the coming year. This panel is intended to share the status of the concurrent efforts and discuss the challenges and potential opportunities associated with eCR with an audience of experienced informaticists.

Objective

Our objectives are to: 1) update the informatics community on the current state of electronic case reporting from healthcare to public health settings; 2) describe informatics methods and tools leveraged for this use case; and 3) share the strengths and limitations of efforts to date and discuss potential challenges and unrealized opportunities.
Panel description

This didactic panel is organized to present differing perspectives on the current status, benefits, and challenges with eCR. The first speaker will provide the vision, describe the need, and share the nuts and bolts of the current status. In addition, the first speaker will share the federal public health view. The second speaker will discuss informatics strategies being used, and the similarities with other efforts, such as the provisioning of knowledge about quality measure reporting requirements. The third speaker will provide a perspective from a healthcare organization based on their participation in several pilot efforts. Finally, the discussion leader is an epidemiologist from a local health department, which is where the ‘rubber meets the road’, and leads the Public Health subject-matter expert workgroup that provides feedback on reporting specifications being developed for the RCKMS project. This experienced team of presenters will be supported by Shu McGarvey, the moderator for this session.

Panel presentation 1: The public health perspective

Laura Conn, Director of the Health Information Strategy Unit in the Center for Surveillance, Epidemiology and Laboratory Services at the CDC, and co-leader of the CDC-funded Reportable Condition Knowledge Management System (RCKMS) project, will provide the big picture of case reporting including the case reporting life cycle from healthcare to state/local jurisdictions, and how it relates to notifiable condition reporting to the CDC. Automating case reporting involves collaboration across public health partners and includes tools, services, and standardized reporting specifications for all reportable conditions. Centralization of services is a focal theme to ease the implementation burden for both jurisdictions and reporters, and part of the electronic case reporting solution includes hosting on the public health community platform (PHCP). The PHCP provides an accessible, flexible, and secure public health information technology platform of interoperable shared solutions governed by and responsive to the public health community. The PHCP will serve to connect clinical EHRs and public health agencies, and integrate RCKMS as a centralized public health decision support tool to determine if a case is reportable and to which jurisdiction(s). Oversight and management of the eCR are key to managing improvements and updates to both tools and content. This presentation includes an overview of the governance strategy and regulatory guidance to support eCR.

Panel presentation 2: Informatics perspective

Dr. Catherine Staes, an informaticist and co-leader of the CDC-funded Reportable Condition Knowledge Management System (RCKMS) project, will describe the informatics-related aspects of the infrastructure presented during the first presentation. For example, the data flow, standards, and knowledge management strategy will be described. The various implementation guides will be ‘pinned to the tail’ on the processes. More importantly though, the components and strategies will be mapped to other efforts currently involving healthcare and public health agencies to illustrate similarities and differences with these other activities. For example, the trigger codes for eCR will be provisioned similar to codes sets required for quality measurement, leveraging experience already implemented for a different MU requirement. Finally, the current status of the eCR components will be shared and potential challenges will be described.

Panel presentation 3: Healthcare perspective

Dr. Shan He, an informaticist with Intermountain Healthcare, will share perspectives on case reporting from a healthcare enterprise. Intermountain has had over 5 years of experience working with early efforts to implement electronic laboratory reporting (ELR) and electronic death reporting to Utah Department of Public Health by developing standard interfaces and integrating clinical decision support (CDS) logic into healthcare workflow. In 2015, Intermountain collaborated with public health authorities, a CDS vendor, and the CDC/CSTE RCKMS team in a pilot project to demonstrate the feasibility of using a centralized public health CDS to replace the current ELR.

As a reporter, Intermountain participated in four areas of work in the pilot: 1) terminology mapping for trigger codes from RCKMS; 2) trigger configuration on the local CDS platform; 3) EHR data extraction for the initial case evaluation payload; and 4) standard payload construction and service connection with the central RCKMS CDS. Trigger code mapping validated the comprehensiveness of the trigger codes defined by RCKMS, even though only a small percentage (2%-15% depending on the condition and the code system) of the set were actively used at Intermountain. Local trigger rules were configured to filter case requests sent to the central CDS service, which complies with the HIPAA minimum necessary requirement and alleviates performance concerns of the central
service. Existing health information technology (HIT) infrastructure including Intermountain terminology lookup and mapping services, decision support services, EHR data extraction services created for HITECH (Health Information Technology for Economic and Clinical Health) certification, and health information exchange (HIE) services were leveraged to the maximum extent in this pilot. Implementation workload will be summarized.

Intermountain also evaluated the effectiveness and timeliness of the condition reporting criteria from RCKMS by analyzing its EHR data. Quality and availability issues of certain data elements used in the reporting criteria were identified. Most laboratory results were recorded in a structured and coded format, whereas diagnoses and symptoms were not always documented in a structured format by physicians. Timely documentation of structured and coded information will improve the completeness and timeliness of electronic case reporting. In the 2015 pilot, an email alert for reportable cases was sent to the designated personnel who will manually send the case report using the specified mechanism. Intermountain is participating the second stage of the pilot in 2016, which supports automated electronic case report generation using EHR data. Findings and lessons learned will also be presented regarding this effort.

**Audience Discussion**

While questions can be addressed after each of the panelist’s presentations, we will reserve 30 minutes at the end for an in-depth discussion with and among the audience. Questions will be posed, such as:

- From the health care enterprise perspective, do the requirements and systems for implementing eCR leverage your existing infrastructure and make case reporting more efficient? Are there unintended consequences that we should be thinking about?
- Do you envision synergy with requirements, systems and processes related to quality measurement and reporting?
- From the informatics perspective, is the current approach leveraging informatics methods and tools to the greatest extent possible? What should be considered for future enhancements? What are the similarities and differences, from your perspective, between phenotyping for research and quality measurement and for case detection for public health reporting. Similarly, what is the same or different between creating a case report for public health and clinical research?
- From the public health perspective, do electronic health records change the thinking about case reporting?

**Conclusion**

Moving to electronic case reporting requires automation of a process that was designed for and conducted by humans. This change requires strong collaboration across public health organizations and eCR activities; a shared mindset toward building robust solutions that support long-term evolution; and governance to support policies, guidelines and security to ease the transition to the new paradigm of eCR.

**References**

Building Software Platforms that Integrate with the EHR: Implementations, Frameworks, and Industry Experiences

Marc Tobias, MD¹, Robert W. Grundmeier, MD¹, Joshua Mandel, MD², Aaron Neinstein, MD³, Janet Campbell⁴

¹The Children’s Hospital of Philadelphia, Philadelphia, PA, USA; ²Harvard Medical School Department of Biomedical Informatics, Boston, MA, USA; ³University of California San Francisco, San Francisco, CA, USA ⁴Epic Systems Corporation, Verona, WI, USA

Abstract
Despite billions of dollars invested nationally and decades of enthusiasm for better interoperability between electronic health record (EHR) systems, successfully integrating custom software applications is not widespread. In some cases, barriers to disseminating these shareable apps arise even when attempting to share with other institutions using the same EHR product. The recent emergence of readily implementable standards such as Fast Healthcare Interoperability Resources (FHIR) and application platforms built upon these standards are rapidly changing the landscape. Additionally, vendor-specific “app-sharing” frameworks are providing new ways for customers to share their developments. These initiatives have taken the form of institutional research vehicles, open standards frameworks, and commercial endeavors. The panelists will describe their experiences developing and implementing platforms that can embed software-like web applications. Example projects will include the use of web services to implement complex decision support rules, shared-decision making projects that directly engage patients, and workflow tools that facilitate communication across health disciplines.

Learning objectives:
1. Describe the current barriers to sharing customized software applications
2. Understand the present-day impact of successful sharing efforts
3. Describe emerging efforts that seek to facilitate the sharing of locally-developed tools

Panel Description
The panelists will describe the motivation and experience with building and implementing software platforms that integrate with the EHR. They will discuss technical and non-technical barriers to their research and implementation efforts, and how they have overcome these barriers to achieve success. Limitations of current technology and future directions, including workflow and usability aspects of app integration, will be presented.

Moderator: Welcome and session overview
Marc Tobias, MD (5 minutes)

Panelist #1: Ten Years of Epic Customization
Robert W. Grundmeier, MD (10 minutes, followed by 5 minutes discussion)
For over ten years, Dr. Grundmeier has been aggressively modifying the Epic Systems EHR to support a wide variety of clinical informatics research efforts. This has been primarily achieved through the Care Assistant, which is The Children’s Hospital of Philadelphia’s (CHOP) internal software integration platform. Unfortunately, despite the local success of these customizations, dissemination beyond CHOP has been a major problem. Dr. Grundmeier will describe how the Care Assistant has been received by clinicians and administration, as well as research that shows
improved patient outcomes through its use. He will also describe the challenges that have hindered dissemination efforts.

Panelist #2: SMART on FHIR
Joshua Mandel, MD (10 minutes, followed by 5 minutes discussion)
Dr. Mandel has helped to foster an ecosystem where third-party health apps can easily connect to data in EHRs, patient portals, and data warehouses. Dr. Mandel will describe the process by which a growing community including FHIR, SMART, and the Argonaut Project is producing open specifications, developer documentation, sample apps, RFP language, and supporting national policy for data access. He will discuss the role of academia, private industry, and government in contributing to improving data access, security, and privacy.

Panelist #3: Cisco-UCSF Connected Health Interoperability Platform
Aaron Neinstein, MD (10 minutes, followed by 5 minutes discussion)
The Cisco-UCSF Connected Health Interoperability Platform (CHIP) leverages emerging interoperability specifications and standards along with Cisco’s broad technology portfolio. Dr. Neinstein will describe how CHIP enables digital health innovators to develop, commercialize, and distribute their applications at scale -- with enterprise-grade features such as security, data virtualization, and data integration. CHIP will facilitate access to EHR data and beyond, including consumer health devices, the Internet of Things, and other new data sources. CHIP will start with an eye toward specific healthcare use cases and workflows that are common pain points.

Panelist #4: EHR Interoperability
Janet Campbell (10 minutes, followed by 5 minutes discussion)
Janet Campbell is a software developer at Epic Systems Corporation. She will specifically discuss the challenges and successes in supporting innovation-sharing among their customers. Supporting issues that arise in customer-authored applications necessitate new approaches, as resolving problems is more difficult when support staff do not have direct access or ability to change source code. System response time may be degraded by poorly written custom software, and the vendor’s role in adjudicating these problems may be unclear. Cultivating a successful and safe application-sharing ecosystem requires new approaches across the software development cycle, from architectural design to end-user support.

Moderator: Facilitated discussion and closing
Marc Tobias, MD (25 minutes)
The facilitated discussion will explore strategies for institutional administrators to offer effective and secure applications within their own EHR. Additionally, the discussion will attempt to provide guidance to software developers regarding how best to approach the current landscape of EHR application integration.

Discussion Questions
The following guiding questions will be used to facilitate discussion:
1. Given what you learned from your projects, what new innovations or changes do you envision in the upcoming years to more meaningfully impact clinical outcomes?
2. How difficult was it to implement tools developed externally within your own health system?
3. What are the biggest barriers to expanding your current platforms?
4. What has been the perception and response to your platform in regards to hospital administration and outside software developers?

Moderator Attestation
The four panelists, Dr. Grundmeier, Dr. Mandel, Dr. Neinstein and Janet Campbell, have all agreed to participate in this didactic panel.
Engaging older adults in the design, implementation and evaluation of Health IT

Anne M. Turner, MD, MLIS, MPH1, George B. Demiris, PhD, FACMI1, Amanda Lazar, PhD2, Blaine Reeder, PhD3
1University of Washington, Seattle, WA; 2Northwestern University, Evanston, IL; 3University of Colorado College of Nursing, Aurora, CO

Abstract
Over the past decade there has been a growing interest in encouraging individuals to take a more active role in their health and health information management. This movement towards greater patient engagement is closely tied to patient centered technologies including patient portals, decision making tools, home based telehealth, sensor technologies and patient access initiatives such as Open Notes. Although older adults make up 18% of the US population and are the largest users of healthcare, their uptake of information technology has traditionally lagged behind that of other age groups. Unfortunately, the needs of the largest group of health care consumers are often not considered because of difficulties involving this age group in research and the misperception that older adults are technology adverse. This panel demonstrates the potential of information technology to improve the health and independence of older adults, and showcases the application of participatory design methods to empower older adults in their health care management and decision making. Panel members will discuss their experience in conducting research with and for older adults in various settings including independent living, retirement communities and assisted living facilities.

Panel Description
The concept of patient engagement refers to greater involvement of patients in plans for health maintenance and clinical decision-making. Patient engagement has been shown to improve health outcomes and patient satisfaction and has the potential to save healthcare costs. Newly emerging technologies such as mobile apps, wearable sensors and patient centered health records are viewed as key components to involving patients in their health care. However, these technologies are often not designed with older adults in mind, despite the fact that individuals over 60 years of age make up 18% of the US population and are the largest consumers of health care dollars. In terms of engagement, this group is often viewed as vulnerable, dependent, and unwilling or unable to try new technologies. As a result their needs are often left out of the design process. How do we encourage participation by older adults in their healthcare and ensure their needs are incorporated into the design of health information tools? Participatory user centered design methods tailored to older adults are being successfully applied in the design of health information technologies including smart homes, information kiosks and personal sensing devices. Each of our panel speakers will discuss their experience developing and using evidence based informatics solutions to help assess and meet the dynamic needs of older adults and their caregivers.

Intended Audience: This panel is intended for designers and researchers who are interested in working with and designing information systems which serve the needs of older adults, their caregivers and other vulnerable populations.

Learning Objectives:
In this session, participants will learn to:
• Discuss approaches to engaging older adults and families and facilitating a shift to patient-centered geriatric care.
• Be able to cite at least four examples of the use of behavioral sensing to identify health issues and provide usable data to elders, their caregivers, and their health care team.
• Identify the major barriers to technology acceptance by elders and ways to overcome these barriers.
• Address factors such as privacy, financial and ethical factors that may complicate conducting research with older adults living in diverse settings and with diverse disabilities.

Panel Organizer and Participants
Anne M. Turner, MD, MLIS, MPH (Panel Organizer)

Dr. Turner is a clinician and public health informatics researcher at the University of Washington, with a primary research interest in how to incorporate the needs and practices of diverse populations into the design of healthcare technologies using user-centered design methods. Dr. Turner oversees the AHRQ funded five-year SOARING project investigating the personal health information management needs and practices of older adults (www.soaringstudy.org). The results of this research will be used to identify design requirements and develop guidelines to better meet the needs of aging populations. Dr. Turner will serve as the moderator for the session.

**Topic Focus:** Panel speakers will discuss their use of participatory design methods to inform the design, implementation and evaluation of health information technologies (HIT) for older adults in the context of a particular project, intervention, or research interest.

George Demiris, PhD

Dr. Demiris is a Professor of Biomedical Informatics at the University of Washington. He will describe the use of “smart home” technologies that facilitate passive monitoring of older adults in their homes with the goal to understand patterns of mobility and overall activity, detect abnormalities and prevent or even predict adverse events with the ultimate goal to support aging in place. He will showcase the potential of behavioral sensing tools to engage patients and their families. Findings from a project based on the deployment of motion and environmental sensors in apartments of 48 community dwelling older adults will be presented. He will also highlight ethical and technical challenges associated with smart home applications as well as opportunities to engage older adults and their families in the design of such systems, as well as the implementation of visual analytics resulting from these large data sets that are meaningful and support clinical decision making. This presentation will focus primarily on ways to engage older adults in the definition of relevant research questions and solicit their feedback, needs and preferences in designing user interfaces that summarize and visualize behavioral sensing data in a way that facilitates understanding and shared decision making.

Amanda Lazar, PhD

Dr. Amanda Lazar is a Postdoctoral Fellow at Northwestern University with research interests focused on older adults with cognitive impairments. Older adults with dementia are often left out of formative research on technologies designed to address psychosocial aspects of dementia. A variety of difficulties with recruitment, logistics, and unclear guidelines regarding ethical issues complicate working with this population. Despite these challenges, it is vital to include the perspectives and experiences of older adults with cognitive impairments in all phases of the technology design and evaluation process. Dr. Lazar will draw on two years of field research in a specific clinical context: art therapy for older adults with dementia. She will discuss findings from her field work within an art therapy program, in which data were gathered through observations of art therapy sessions, interviews with art therapists, and the design and evaluation of new technologies to support sharing. Dr. Lazar will discuss the potential benefits of new technologies that support sharing therapeutic artwork and tensions inherent in encouraging sharing within a protected health context. She will highlight issues that arose regarding ethics, privacy, and empowerment. This presentation will focus primarily on issues related to ethics, privacy, and empowerment that arise when conducting participatory research with individuals with cognitive impairments.

Blaine Reeder, PhD

Dr. Reeder is Assistant Professor of Health Care Informatics in the College of Nursing, University of Colorado. He will discuss the differences in usability and usefulness of smart watches and smart home sensors for health-related data collection. Dr. Reeder will present the results of a technology acceptability study that enrolled 10 older adult women (young-old and middle-old) for interviews that compared perceptions of wearable activity monitors and home-installed motion sensors. In addition, he will present an overview of current Pebble smart watch technology with a focus on smart watch potential to facilitate participatory design research with older adults. Smart watch features that improve upon current research-grade activity monitors will be highlighted, including: built-in design for translation based on consumer-grade device availability; real-time access for physical activity data; wrist-based mini-surveys; bi-
directional communication with researchers and providers; and older adult cohort familiarity with watches as a wearable technology that enhances acceptability and eliminates medical device stigma. Lessons learned from smart home and wearable device studies that engaged older adults and the potential for smart watch integration with Internet of Things platforms will be presented as takeaways. This presentation will focus primarily on how perceptions of wearable and home-installed sensors can inform participatory design of technologies for older adults.

Participation
All participants have agreed to take part in this panel.

Discussion Questions:
- How do we engage family caregivers and other stakeholders in the design, implementation, and evaluation of health IT?
- How do attitudes and values of elders differ from those of family members and caregivers?
- Can design features overcome stigma associated with being older?
- What are some techniques for reaching out to older adults with cognitive and physical limitations in participatory research?
- How do we encourage technology designers and commercial vendors to consider older adults in the design of HIT?
AMIA Fall Symposium 2016

Panel Proposal

Transforming Clinical Documentation: Nursing and Big Data Working Group 10’s Recommendations and Action Plan

Organizer: Charlotte A. Weaver, RN, PhD, FAAN

Panelists
Charlotte A. Weaver, RN, PhD, FAAN, AMIA Co-Chair to ANI, (caweaver2011@gmail.com)
Judith Effken, PhD, RN, FACMI, FAAN, Professor Emerita, University of Arizona, College of Nursing (jeffken@nursing.arizona.edu)
Kelly Cochran, MS, RN, Policy Advisor, HIT, American Nurses Association, (kelly.cochran@ana.org)
Rebecca Freeman, PhD, RN, PMP, Chief Nursing Officer, Office of the National Coordinator, (Rebecca.freeman@hhs.gov)
Ann O’Brien, MSN, RN, CPHIMS, Sr. Director Clinical Informatics, Kaiser Permanente, (ann.o’brien@kp.org)

Abstract
In acute care, nursing documentation in today’s electronic health records systems (EHR) has become so burdensome that its sheer volume ensures that virtually no member of the care team reads it, not even nurses. EHR documentation was created in the image of the paper chart with a focus on data capture. Consequently, data views that could assist clinicians in assessment, on-going evaluation, and team communication/data sharing are mostly woefully absent. Thus, instead of promoting nurses’ core functions of monitoring and early-warning surveillance, this data-rich but information-poor (DRIP) state undermines patient safety by not supporting effective sharing of information among nurses or with the care team. This panel will describe how nurses view current EHR systems and how they restrict nurses’ ability to provide optimal, safe, efficient, and quality patient care. Panel members will present current initiatives addressing some of these pain points and offer specific recommendations for how the best-practices resulting from these initiatives might be rapidly shared with others through a vendor-neutral repository.

Learning Objectives: After attending this panel presentation, attendees will be able to:
1. Describe the current documentation burden built into today’s EHRs for nurses.
2. Discuss academic research on EHR usability, design, and nurse satisfaction with EHRs.
3. Describe 3 current initiatives by nurse leaders for transformative change in clinical information systems design.
4. Describe the recommendations and action plan of the Nursing and Big Data’s Working Group 10.

Description
In 2014, a group of informatics and nurse leaders from practice, academia, and the vendor community formed because of a growing recognition that current EHR systems were not supporting evidence-based, nursing practice and lacked the key functionality needed for clinical decision support and safe patient care. The recommendations and action plan that will be presented by this Panel draw upon two years of work by the Working Group that emerged from the initiative. AMIA’s EHR 2020 report aptly described the limitations and vulnerabilities of current state EHRs for supporting safe, effective and efficient care, but focused on the documentation burden imposed on physicians. In 2014, the nurse leaders attending the University of Minnesota’s “Nursing and Big Data” conference issued a similar call to action. Underlying this call was the recognition that nurses’ issues about the challenges of EHRs to meet their needs was rapidly reaching a tipping point. The experiential evidence reported by conference attendees has been corroborated by recent studies reporting poor nurse satisfaction scores related to EHR documentation requirements and excessive time needed to complete admission assessments and a demand that documentation support nurses’ practice and improved patient outcomes rather than prioritizing the legal, regulatory, reimbursement needs of the organization. Nurse practice leaders are expressing concern that at some point, nurses will organize around and become increasingly vocal about their EHR dissatisfaction.
Current EHRs are not designed to guide the delivery of highly reliable, evidence-based nursing or team care practice\textsuperscript{4,6,9}. Even when a health care organization purchases an EHR system, they must design, build and implement it without the benefit of best practices benchmarks. There are few mechanisms available to leverage lessons learned from other organizations or to access a “best practice” central repository that holds “how to” examples, such as data sets complete with clinical terms mapped to the standardized terminologies of LOINC and SNOMED-CT. Instead, each health care organization starts from their vendor’s basic system and must depend on their singular-threaded expertise. Often, the clinical personnel tapped to design and build the new clinical system are nurses, physicians, and health professionals who show an interest in the project but have no formal informatics education or certification. A major theme for nurses is that their documentation burden has steadily increased because of the addition of new regulatory mandates, quality report requirements, or overzealous interpretation of these requirements\textsuperscript{6,8,10-11}. Unfortunately, as new requirements are added nothing is removed. One nurse informatics leader described it like a home that had been lived in through all the family life cycle, with nothing ever given away and every room filled—including the attic\textsuperscript{10}. A drastic change is needed in the way we design and build our EHRs so that we can benefit from lessons learned. We also need system functionality that provides the following: multiple technologies for documenting care; support of nurses’ workflow with real-time decision support; and, flexible data visualization and screen presentations that can be specific to and changed by the clinician. Added to these needs are enhancements that allow nurses to capture and share the patient’s story clearly and concisely so that team communication is optimized.

In acute care organizations, nurses represent the largest users of health information technology and are responsible for more documentation than any other group of health professionals\textsuperscript{12-13}. For that reason, if goals for transforming our healthcare systems’ quality, safety and costs are to be achieved, nursing has to be included and given tools that optimally support these care delivery goals. Currently, there is little understanding of how EHRs should be built to support nurses’ ability to deliver safe, evidence-based care. Thus, our Working Group’s first year report aimed to share the available evidence broadly among nursing leadership, and was published in the Nursing Administration Quarterly rather than in an informatics journal\textsuperscript{6}. This 2015 report outlines the problems with the EHR, and also sets out principles and specific recommendations for optimizing EHRs to improve the processes of care delivery and generate actionable information for nursing care decisions and quality outcomes improvement. Now we are bringing our work into health policy and informatics domains\textsuperscript{14}. This Panel will present a specific set of go-forward actions detailing the starting content for a vendor-neutral best practices repository, possible hosting sites, and a plan for assuring the repository’s maintenance and sustainability over time.

References:

Panel Organization:
Intro/Scope/Problem  Weaver  15 minutes
Academic Work      Effken  15 minutes
Policy             Cochran  15 minutes
Exemplar          Freeman  15 minutes
Q & A             O’Brien  30 minutes

Panelist Roles:
Ann O’Brien, MSN, RN, CPHIMS. Ms O’Brien, Working Group 10 Chair, will moderate the panel discussion and audience questions.

Charlotte A Weaver, PhD, RN, MSPH, FAAN. Dr Weaver (panel organizer) Working Team Chair will present the scope of work completed, the problem being addressed and the intended approach to identifying best practices for designing optimum, evidence-based documentation and exemplars mapping nursing problems to LOINC and SNOMED-CT, including those resources available at NLM and on other web sites.

Judith Effken, PhD, RN, FAAN, FACMI. Dr Effken, Working Team Co-chair, will highlight the major academic initiatives informing the recommendations, approach and content for a Library of Best Practices demonstration project.

Kelly Cochran, MSN, RN. Ms Cochran will highlight the policy perspectives that support this initiative to transform clinical documentation with the driving priority being improving patient safety and quality while cutting costs. The Washington DC landscape and multiple entities that represent stakeholders and supporters will be discussed.

Rebecca Freeman, Ph.D, RN, PMP. Ms Freeman will present a vendor-neutral technical specification approach for streamlining nursing documentation developed at a large, national healthcare system using multiple EHR vendors.

Participation Statement
All proposed panelists are aware of this panel submission and have agreed to participate in the panel if the proposal is accepted.

This Panel is a NIWG sponsored submission
Informatics to Transform Med Wreck to Medication Reconciliation

Mark G. Weiner, MD, Charlene Weir, PhD, RN, Terrence J Adam, PhD, MD, Edgar Chou, MD

1Department of Clinical Sciences, Temple University School of Medicine, Philadelphia, PA
2Department of Biomedical Informatics, University Of Utah, Salt Lake City, UT
3Institute for Health Informatics, University of Minnesota, Minneapolis, MN
4Division of Internal Medicine, Drexel University College of Medicine, Philadelphia, PA

Abstract: Appropriate conduct of Medication Reconciliation is a fundamental component of meaningful use, and an important contributor to patient safety. In the ambulatory setting, medication reconciliation aims to ensure that the patient and provider are in synch with the medication regimen. In the inpatient setting, the process mandates that medications from a prior treatment location are appropriately addressed in the transition from one level or location of care to another. At discharge, medication reconciliation requires thoughtful decisions about continuing inpatient medications into the ambulatory setting, with appropriate recognition of prior outpatient medications that should be stopped or changed. This panel will discuss some of the informatics, workflow, communication, and human challenges that interfere with optimal completion of medication reconciliation and several institutional initiatives designed to address these challenges.

Learning Objectives:
1. To develop a systematic approach to identifying and addressing problems in medication reconciliation
2. To understand the patient perspective in managing medication reconciliation
3. To identify the implicit cognitive goals involved in medication reconciliation and their implication for display designs.
4. To explore informatics solutions that ease the process of conducting medication reconciliation for clinicians and patients.

Documentation of pharmacologic orders in the inpatient and ambulatory settings has been greatly improved in the era of Electronic Health Records. While some ambulatory scripts are still written on old paper scripts, most medication ordering takes place through the EHR, creating the unprecedented ability to track all medications that have been ordered for a patient. However, the simple availability of these medication orders does not always translate into a simplified medication reconciliation process that successfully aligns active medication lists in the EHR with what the patient has been taking, and what the patient SHOULD be taking. While medication reconciliation is required at all ambulatory visits, and at all transitions of care in the inpatient setting, there are a number of factors that make this process more difficult than is commonly realized. Medications can still “fall off” the medication list. Patients can still take medications in a manner differently than prescribed. Most pharmacies are still not receiving orders to discontinue medications, so patients often pick up refills of “old” medications, while also receiving new prescriptions that were intended to replace the old medications. In addition, medication reconciliation results in significant cognitive demands on the part of the provider and unnecessarily breaks up the flow of care. Informatics tools can present a more complete picture of medications that have been prescribed from other providers, and can also show medications that the patient has retrieved. If properly designed and operationalized, the use of informatics solutions can reduce cognitive burdens on provides and facilitate a higher quality medication experience for patients.

This panel will explore medication reconciliation at 4 institutions, and discuss the human, workflow and technology issues faced and solutions being implemented in each setting.
Mark Weiner, MD, Temple University Health System

Temple University Health System currently runs Epic Ambulatory and is the final phase of implementation of the Inpatient product. The Ambulatory workflow for medication reconciliation begins with Medical Assistants who asks the patients about each medication in their medication list and whether or not they are taking or not taking it. However, often times they are not sure, or even if they carry a medication list they have received elsewhere, they still may not be sure what they are taking. Through the course of the visit, the physician has access to data on meds picked up through Surescripts, and can also link to medication lists at other sites through Epic Everywhere. This establishes a discussion of what medications the patient may be taking, though a great deal of ambiguity exists. In the primary care setting, the broad range of health issues addressed often allows for thoughtful decisions about revising the medication list, though in specialty settings, the focus of the visit may not allow reconciliation of medications outside the domain of the clinical specialty. When medications that the patient has stopped on their own are removed from the medication list, the after visit summary associated with the visit informs the patient to stop the medication, making it appear that the provider had recommended the discontinuation at the time of the visit. Medications that get renewed are given a start-end date interval associated with the interval from the prior instance of that medication which, if it is too short, may make the medication fall off the list much sooner than expected. Dr. Weiner will outline and categorize these and other pitfalls in the medication reconciliation process.

Charlene Weir, PhD, RN, University of Utah

The VA currently has integrated a number of tools and policies designed to ensure continuity of medication treatments, including note templates, policies for discharge follow-up, software that identifies refill rates and the use of pharmacists in the inpatient and outpatient care settings. Across a series of ethnographic observations, cognitive task analyses, focus groups and interviews, VA researchers have identified some key barriers to the design and implementation of informatics solutions for effective medication reconciliation. The first is that the “task” and meaning of medication reconciliation varies always across roles, with each role bringing different goals and perspectives to the process. Med Rec can range from a simple matching task, to a deep assessment of the appropriateness of treatment across patient and disease. The implication is that medication reconciliation policies must identify who’s responsible and informatics solutions need flexibility in display. The third issue is that medication reconciliation is often viewed as an assessment task alone, when often it is associated with decisions and actions that need support. Finally, the task of medication reconciliation at times is disembodied from the patient-provider interaction because it is not part of the natural workflow. Dr. Weir will describe assessment, workflow and logistical requirements to enhance the medication reconciliation process.

Terrence J Adam, PhD, MD, University of Minnesota

The care transition process has the potential to create patient safety concerns due to problems of communications, logistics and care complexity issues. The care transitions create a need for the health system to effectively understand the patient’s condition and pass pertinent information from care facilities to other care facilities, patients and caregivers. A particular transition of care problem is patient communication, verification and management of medication regimens. The acute care setting can create a number of confusing messages for patients who are experiencing illness and may hear from multiple specialists during their care experience.

To address this problem, a structured data approach based on the pharmacist drug therapy problem framework can help identify pertinent medication issues and help provide guidance for their resolution. The University of Minnesota and affiliated Fairview health system have used EMR driven business process rules to identify at-risk patients and facilitate transition of care management of medication regimens. By using a mix of information management solutions and care process follow-up, the University-affiliated health system has created a mechanism to systematically identify patients and intervene with pharmacist review using in-person and phone based follow-up.
Dr. Adam will discuss the role of informatics in supporting an improved medication reconciliation process, with a special focus on care transitions.

**Edgar Chou, MD  Drexel University**

Drexel University College of Medicine began the journey of its ambulatory electronic health record implementation, Allscripts Touchworks, in 2006. Drexel is the academic affiliate with its partner hospital, Hahnemann University Hospital, which has its separate electronic health record, Cerner. Similar to most multispecialty practices, challenges are abound in ensuring the medication list is adequately reconciled, ranging from patient’s literacy to the numerous players that now impact whether a medication is currently being taken or not or can continue to be taken or not for the primary care physician and for the medication list to be truly “reconciled” - these extend beyond the specialist to include the insurance payer (changes in formularies and coverages and participation in Surescripts), the pharmacy benefit plan, and outside institutions (hospitals, skilled nursing, rehab) and agencies (home health) that are involved in the care of the patient. Informatics solutions have addressed challenges in the electronic prescribing process over these years. We now have the ability to initiate prescriptions to pharmacies, and receive information on filled prescriptions that providers can verify before importing into our EHR. Providers receive prescription renewal requests automatically through the EHR, and see drug formulary reminders at the time of placing prescriptions or renewals. Lastly an increasing number of providers have the dual factor authentication required to prescribe controlled substances directly to pharmacies across all 50 states. Despite these technological advancements, time has still been the greatest barrier in achieving this across our institution. We have embraced this reality with incorporating additional resources to assist with the process. We have expanded certified Medical Assistants’ roles and incorporated nursing staff to assist with this process. Dr. Chou will describe how informatics solutions have been incorporated into clinical workflows to improve medication reconciliation.

**Intended Audience and Timeliness**

The intended audience includes physician and nurse informaticists as well as academic informaticians interested in the understanding trends and limitations in technology designed to foster medication reconciliation and the continued need for workflow improvements and trained people in the process. This is an especially timely topic as improved interoperability of information systems and the associated increased access to medication data from pharmacies, payers, and other institutions provides better information substrate for medication reconciliation, and more opportunities for recognition of important medication conflicts. In addition to robust technology, true medication reconciliation requires a team approach to fully address the detail from multiple information sources and the complexity of many patients’ medication regimens.

All participants have agreed to take part on the panel.  

This submission is endorsed by the Primary Care Informatics Working Group

**Discussion Questions**

- What are the unique information needs in managing medications through transitions of care?
- How does interoperability both help and add to the complexity of conducting medication reconciliation?
- How can the drug therapy problem framework help in systematically identify at-risk patients?
- Does the use of drug therapy problems help improve post-discharge patient follow-up?
- How do institutions organize teams to manage complex medication regimens?
- What are the remaining informatics and workflow challenges in medication reconciliation?
Abstract

Often the focus of public health informatics is narrowed down to application of IT systems and other technology solutions in public health settings. We argue that actually that scope should be broadened, first, to analysis and improvement of business practices utilized by public health agency program; and second, to a much wider application of information and computer science technology to promote and improve the public’s health, well beyond the reach of public health agency programs. This panel will provide a variety of perspectives on use of business analysis techniques for typical public health informatics projects.

Learning objectives. After participating in this session, the participant should be better able to:

- Formulate an approach to application of business analysis techniques to document institutional knowledge, analyze and improve performance of key functional aspects of public health programs, and develop requirements for technology solutions.
- Reduce complexity of a typically complex public health program by using business analysis to break it into component parts to explore how parts work together to create an efficient whole.
- Replicate the approach and apply lessons learned on how methods of business analysis have been used to develop operational best practice guidelines for immunization information systems and in other settings.

Panel Description

This panel will provide a variety of perspectives on use of business analysis techniques for typical public health informatics projects. Often the focus of public health informatics is narrowed down to application of IT systems and other technology solutions in public health settings. We argue that actually that scope should be broadened, first, to analysis and improvement of business practices utilized by public health agency program; and second, to a much wider application of information and computer science technology to promote and improve the public’s health, well beyond the reach of public health agency programs.

Issues that will be examined:

- Role of the business analysis component in public health informatics.
- Using business analysis techniques to analyze and improve public health operations and processes.
- Using business analysis techniques for clinical decision support for immunizations.
- Other examples of using business analysis techniques to improve public health practice and outcomes.

Moderator: Warren Williams, MPH

Brief Description of Presentation 1: Role of the business analysis component in public health informatics. Patrick O’Carroll, MD, MPH.

Public health informatics is primarily an engineering discipline. Engineering techniques of business analysis should be applied to understand the business of public health – to elucidate in concrete terms exactly what the public health agency does. This step, seemingly straightforward, can in fact be the most difficult and time-consuming element. It involves the development of models of the business by use of formal modeling techniques.

Brief Description of Presentation 2: Using business analysis techniques to analyze and improve public health operations and processes. David Lyalin, PhD.
Business analysis and facilitation techniques, as well as consensus building methodologies were used to explore immunization information systems (IIS) processes and operations. IIS experts analyzed existing practices, collaborated to identify process details, evaluated and combined various operational practices, and achieved a consensus regarding best practice recommendations. These best practice recommendations offer practical guidelines on the most challenging operational areas for IIS, such as data quality assurance, reminder-recall, deduplication of immunization records, patients eligibility for public programs, vaccine inventory management, coverage assessments. Also, developed guidelines serve as a source of requirements for IT solutions, illustrating the notion that operational needs should drive technological solutions, not the other way around.

A number of business modeling techniques have been used to capture best practice recommendations. For example, we have applied domain modeling in a variety of public health areas, providing a common vocabulary and establishing a foundation for other model types. Decision models, such as business rules and decision tables, were utilized to unambiguously document high-level policies, institutional knowledge, and operational-level decision-making. Process models, such as use cases (structured description of operational scenarios) and a variety of process diagrams, have been employed to describe processes and process participants. Event models helped us to analyze events that lead to change of statuses for various public health concepts (e.g., status of a vaccine dose during its life cycle). These analysis models supported three stages of our collaborative work: discovery (where the stakeholders documented understandings of how exactly the current program operations work, which sometimes isn’t clear to everyone, in all areas of operations), assessment (i.e., what is working well and what isn’t, as well as discussions of improvement options), and specification (i.e., agreed-upon program requirements that should be implemented).

Independent evaluation findings indicate that application of these best practice recommendations in the IIS domain of state and local health departments resulted in improved data quality, reduced staff time, and increased efficiencies across immunization programs. Vendors are using these guidelines as a reference source to implement operational best practices a variety of IT platforms.

Brief Description of Presentation 3: Using business analysis techniques for clinical decision support for immunizations. Stuart Myerburg, JD.

Immunization clinical decision support (CDS) is an automated process that occurs within Immunization Information Systems (IIS) and Electronic Health Records (EHRs) to determine the recommended immunizations needed for a patient and deliver these recommendations to the healthcare provider. These recommendations are developed by the Advisory Committee on Immunization Practices (ACIP). After ACIP recommendations were published, technical and clinical subject matter experts (SMEs) worked to interpret and integrate them into their CDS engines. The translation of that clinical language into technical logic was time-consuming, complex, and done mostly independently within each system. As a result, CDS engine outputs often varied and did not always match the expectations of clinical SMEs.

The Clinical Decision Support for Immunization (CDSi) project was begun to provide a single, authoritative, implementation-neutral foundation for development and maintenance of CDS engines. Business modeling techniques were used to develop the CDSi resources. Domain and process models, along with decision tables and business rules, were used to capture ACIP recommendations in an unambiguous manner and improve both the uniform representation of vaccine decision guidelines, as well as the ability to automate vaccine evaluation and forecasting.

The target audience for the CDSi resources includes business and/or technical implementers of immunization CDS engines. CDSi provides them with a bridge between the clinical recommendations and the rigorous technical logic needed for their systems. The business modeling used in CDSi provides uniformity and reduces ambiguity, assuring patients receive the right immunization at the right time.

Brief Description of Presentation 4: Examples of using business analysis techniques to improve public health practice and outcomes. Bill Brand, MPH.

This final section will highlight several examples of using business process analysis and re-design to improve practice and efficiency. The examples will include automating previously manual workflows, designing public health processes to better match clinical workflows in healthcare settings, innovative approaches and design in
surveillance, and development of shared services to meet common needs across organizations. It will close the didactic portion of the session with overall principles and guidance that distill the need for and benefits of collaborative approaches to business analysis in public health as it seeks to modernize its processes and information systems.

**Why the topic of this panel is timely, urgent, and needed**

Implementation of formal scientific methods of statistics for data analysis defines public health as an applied science and provides a basis for policy and operational decision-making. At the same time, systematic analysis of operations and processes where information originates has limited implementation in the public health enterprise. Methods for such business analysis have been developed within a field of industrial and systems engineering. Application of these methods helps to document institutional knowledge, analyze and improve performance of key functional aspects of public health programs, and develop requirements for technology solutions.

As complexity of the public health enterprise increases, systematic analysis techniques help to deal with that complexity. Specifically, business analysis allows to break a typically complex public health program into its component parts, i.e., common perspectives of why, who, how, where, what, when. Such a partition makes it easier to both understand the program (from its policies to technical implementations) and to explore how its parts work together to create an efficient whole. Benefits of using the business analysis are illustrated with case studies from several public health informatics projects.

Anticipated audience includes public health informatics professionals from government, private sector, and academia.

**List of discussion questions:**

- What more can be done to educate public health students and practitioners in informatics principles and approaches, such as business analysis?
- How can business analysis techniques enable public health and healthcare organizations to better understand each other’s workflows and business requirements? In what areas could we collectively do more to ensure synergistic processes?

**A statement from the panel organizer:**

All participants have agreed to take part on the panel.
Mental Health/Substance Use Disorder Care, Privacy, and HIT—Can We Make It All Work?

David W. Bates, MD1, Alisa B. Busch, MD, MS2,3, Marissa Gordon-Nguyen, JD, MPH4, Paul C. Tang, MD, MS5

1Brigham and Women’s Hospital, Boston, MA; 2McLean Hospital, Belmont, MA; 3Department of Health Care Policy, Harvard Medical School, Boston, MA; 4US Department of Health and Human Services, Office for Civil Rights, Washington, D.C.; 5Palo Alto Medical Foundation, Palo Alto, CA

Abstract

Electronic health records (EHRs) enable easier, more complete sharing of health information across a patient’s providers. Mental health and substance use disorders (MH/SUDs) are common, and often cause or exacerbate other medical conditions, so patients could benefit from integration of MH/SUD information into EHRs and other health information technology. However, patients view some health conditions as more private than others and would like more control over how, or to whom, their health information is shared. For example, patients expect, and some federal and state regulators require, special privacy protections for MH/SUD care, which can pose challenges for EHRs and health information exchange (HIE). Overcoming the challenges of applying additional privacy protections to specific data categories in EHRs and HIE has important implications for patient autonomy, care-seeking, and healthcare quality and safety. It also has important implications for behavioral health providers’ ability to participate in HIE, and for our healthcare system to move towards increasingly efficient, higher quality, and safer healthcare. Recent federal and state efforts are aimed at addressing some of these issues. The goal of this interactive panel is to discuss the ethical, medical, public health, regulatory and technological challenges and controversies of addressing privacy in EHRs and HIE.

Intended Audience: Clinicians, informaticists, policy makers

Introduction

There are important technological, political, and clinical challenges to meeting the goals of respecting patient autonomy in information sharing using electronic health records (EHR) and health information exchange (HIE), while also meeting the medical and public health goals of data sharing and a robust learning medical system. A 2014 Office of the National Coordinator (ONC) survey finds that while the proportion of individuals who worry about the privacy of their healthcare records declined recently, a substantial minority (30%) remain concerned, and 5% reported withholding information due to their concerns. Similarly, studies in primary care samples have found that about one-third would restrict access of sensitive information to at least one healthcare provider in their EHR. These statistics reflect privacy concerns broadly and for a variety of sensitive medical conditions. Mental health and substance use disorders (MH/SUDs) represent illnesses for which patients often have heightened privacy concerns related to discrimination and stigma, at times from their healthcare providers.

As a result, federal and state policies regulating the disclosure of MH/SUD records are often more stringent than for healthcare in general, even for treatment purposes. The MH and SUD privacy regulatory environments differ somewhat, but both pose challenges to EHR implementation and HIE. For mental health, the federal Health Insurance Portability and Accountability Act (HIPAA) privacy regulations protect the privacy of individually identifiable health information and generally does not distinguish between information about mental health and other types of health information. However, states commonly have mental health privacy laws that augment HIPAA. In contrast, the federal SUD privacy regulation, 42 CFR Part 2, rarely is augmented by state law. Proponents of integrated, HIT-enabled care argue that 42 CFR Part 2 is outdated and contains requirements that are
extremely difficult to implement in HIT.\textsuperscript{9,10} Others argue that its protections are necessary to avoid discrimination in personal and patient care experiences.\textsuperscript{11}

Recognizing the importance of patients with MH/SUDs being able to benefit from HIT (including healthcare delivery integration and an enhanced ability to assess and improve healthcare quality), the importance of respecting patient autonomy for consent, and the unique medical record privacy regulatory environment of MH/SUD care, there are a variety of active federal efforts to address these issues. For example, the ONC and the Substance Abuse and Mental Health Services Administration (SAMHSA) are working to address HIE technological challenges of data segmentation and consent.\textsuperscript{12,13} And in recognition of the challenges posed by 42 CFR Part 2, in February 2016 the SAMHSA released proposed changes for public comment.\textsuperscript{14}

Despite the gains from these efforts, much policy and technological work remains. For example, data segmentation was not deemed ready enough to be required in the October 2015 EHR Certification Final Rule\textsuperscript{15}. And, while the proposed rule change to 42 CFR Part 2 has provisions for HIT environments, it is unclear whether the changes will adequately both address public concerns about discrimination, and dismantle the barriers to EHR implementation and HIE in MH/SUD care. Also, patient restricting authorization for clinical information sharing across providers can have important adverse consequences for the quality and safety of healthcare. Emerging evidence finds that developing the appropriate technology and privacy policies will be inadequate absent robust patient education about the implications of their choices.\textsuperscript{2}

This panel topic is both timely and controversial because of the active federal and state efforts to address MH/SUD data segmentation and privacy, support HIE of MH/SUD treatment, and revise a controversial federal policy (42 CFR Part 2) which impedes HIT, HIE and integrated healthcare for individuals with SUD but is also seen by some as providing needed protection against healthcare discrimination. These efforts are in the context of an emerging but not quite ready technology and policy landscape to enable granular consent of information, nor a clear understanding of what patients will need to support them in making educated decisions about information sharing. Also, these efforts are in a healthcare policy context in which alternative models of payment and healthcare delivery are making coordinated and integrated general medical and MH/SUD care an increasing imperative and reality—often absent MH/SUD information about patients that could have important implications for care.

Panel Aim and Expected Discussion: The aim of this interactive panel is to discuss the ethical, medical, public health, technical and policy/regulatory challenges of HIT privacy in the context of MH/SUD care, and the changing practice models of integrated, coordinated healthcare.

Expected Contribution of Each Speaker: All panelists have agreed to participate in this panel. **David W. Bates, MD** is a general medical internist, the Chief Innovation Officer at Brigham and Women’s Hospital, and an expert in quality and safety in electronic medical records, and health information technology policy. Dr. Bates will serve as moderator and also discuss how mental health issues and substance use are actually being managed in one large integrated delivery system, including the challenges of data segmentation and anonymization. There is considerable variability among both systems and states in terms of how data around MH and SUD are both represented and especially transferred. **Alisa Busch, MD, MS** is a psychiatrist, the Chief Medical Information Officer, and Director of Clinical Performance Measurement and Health Services Research at McLean Hospital. She is also an attending psychiatrist in the McLean Hospital Alcohol and Drug Abuse Treatment Program. Dr. Busch will discuss the state of HIT implementation in MH/SUD care, the privacy regulatory landscape unique to MH/SUD programs and implications for data sharing within EHRs, and federal and state efforts to facilitate EHR implementation and HIE in MH/SUD providers. **Marissa Gordon-Nguyen, JD, MPH** is a Senior Health Information Privacy Policy Specialist at the US Department of Health and Human Services, Office for Civil Rights, Washington, D.C. Ms. Gordon-Nguyen will discuss the HIPAA provisions that address quality assessment/improvement and care coordination, as well as guidance materials related to mental health disclosures. **Paul C. Tang, MD, MS** is the Vice President, Chief Innovation and Technology Officer at the Palo Alto Medical Foundation, and Vice Chair of the Health IT Policy Committee for the Office of the National Coordinator for Health Information Technology Committee (HITPC). Dr. Tang will discuss HITPC recommendations on sensitive health information, including Part 2 information.
References:


New Pathways Into Biomedical Informatics: Educational Outreach Programs for High School Students

Panelists: David Boone, PhD1, John T. Finnell, MD2,3, Kim M. Unertl, PhD4
Moderator: Indra Neil Sarkar, PhD, MLIS5

1Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; 2Regenstrief Institute, Indianapolis, IN; 3Department of Emergency Medicine, Indiana University School of Medicine, Indianapolis, IN; 4Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; 5Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

Biomedical informatics education in the United States started with a focus on the graduate and post-graduate level. As the field has evolved, educational opportunities have expanded into new areas, such as certificate and undergraduate programs. More recently, biomedical informatics programs have begun exploring a new frontier for the field: educational outreach to high school students and high school teachers. Questions remain about these outreach efforts. At a foundational level, why is educational outreach to high school students important to biomedical informatics? At a more pragmatic level, what types of structures are necessary to support meaningful involvement of high school students in research? Panelists representing different models of high school outreach will give an overview of program design at their institutions. The panel will then collectively discuss what has gone well in their programs, along with the barriers or challenges they have encountered. The panel will have an interactive format to enable a robust discussion around both foundational and pragmatic questions about the role of educational outreach to the high school level in the future of biomedical informatics.

Introduction

Biomedical informatics is a rapidly evolving field, with growing needs for a wide range of interdisciplinary skillsets. The field is also rapidly expanding with the advent of Meaningful Use, Precision Medicine, and other major initiatives focused on applying computational approaches to health and healthcare. Biomedical informatics education originated with an emphasis on graduate and post-graduate education; as the field has grown, the need and opportunity for outreach to broader educational levels has also grown. Several biomedical informatics educational programs have explored the idea of developing undergraduate education programs, including formal degree programs, through approaches that include both didactic coursework and research experiences1,2. Looking to engage students even earlier in their educational trajectories, several programs have also explored outreach to the high school level, through a variety of different approaches. Little is documented in the literature about the scope of high school outreach efforts, the impact of high school outreach on students and researchers, and what components are needed for successful and meaningful outreach to the high school level.

Since its inception, the American Medical Informatics Association (AMIA) Annual Symposium has supported dissemination of research conducted by students at the graduate and undergraduate level through programs such as the Student Paper Competition. To support dissemination of research conducted by high school students working with biomedical informatics researchers, the AMIA High School Scholars Program (HSSP) was introduced at the 2014 AMIA Annual Symposium3. The AMIA HSSP, guided from its inception by Finnell, Sarkar, and Unertl, provides an opportunity for selected high school students to present their research and participate in a major scientific meeting. Over the past several years of working with the AMIA HSSP, we have received many questions about how to get started with outreach to the high school level, the benefits and challenges of working with high school students, and the role of high school outreach in the future of the field. Despite substantial support for the AMIA HSSP, we have also repeatedly encountered the perception that engaging students earlier in their education presents significant challenges to potential mentors and to biomedical informatics educational programs.

Aim of the panel discussion

The idea of engaging high school students in Biomedical Informatics has seen substantial support within AMIA and in biomedical informatics programs at several universities. Many questions remain, however, about how to get
started with outreach to the high school level, what types of elements are needed for a successful engagement with high school students, and effective approaches for researchers seeking to work with high school students. Our panel aims to answer these questions based on our extensive experiences with high school outreach and to have a discussion with the audience regarding perspectives of audience members on high school outreach.

As the AMIA HSSP enters its third year, the time is ripe to engage the AMIA community in a discussion about:

1. Why it is important to expand biomedical informatics outreach to the high school level,
2. Potential roles for biomedical informatics in advancing STEM-C (Science, Technology, Engineering, and Math, including Computing) education in the United States, and
3. How individual researchers and organizations can overcome barriers to working with high school students.

The interactive panel format will allow the panel and the audience to engage in a robust discussion around these key issues and concerns, and also potentially to contribute to future directions of the AMIA HSSP.

**Description of panelist contributions**

The panel brings together panelists representing different approaches to high school outreach and a moderator with extensive experience with the AMIA HSSP.

**David Boone** is an Assistant Professor in the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh with research interests in long noncoding RNAs, breast cancer biology, and cancer transcriptomics, in addition to a keen interest in education and outreach. He is the Director of the Computer Science, Biology and Biomedical Informatics (CoSBBI) high school summer research program and the new Internship in Biomedical Research, Informatics, and Computer Science (iBRIC) for undergraduates. CoSBBI was founded in 2011 by Michael Becich, the chair of the DBMI at the University of Pittsburgh. The CoSBBI program brings a diverse group of high school students to Pittsburgh for 8-weeks of authentic and mentored research and career preparatory experiences. Students are matched with research mentors, who direct the scholars’ work on independent research projects aligned with the interests of the mentor. Additionally, students are exposed to workshops, seminars, journal clubs, and research roundtables that teach students research and career skills and expose them to various biomedical informatics topics. To perpetuate the pipeline, students that successfully complete CoSBBI are invited to apply for a paid summer research internship as part of the iBRIC program for the duration of their undergraduate education. The program continues to expand, with 10 CoSBBI scholars and 6 interns participating in the summer of 2015 and plans for 10 CoSBBI and 20 iBRIC students for the upcoming summer. CoSBBI is funded by the NIH, Doris Duke Charitable Foundation, University of Pittsburgh Cancer Institute, University of Pittsburgh Medical Center, and other charitable and foundation giving as part of the UPCI Academy, an outreach effort that spans multiple departments at the University. David is the executive director of the UPCI Academy. iBRIC is funded by the DBMI, NIH, PA-CURE, and other charitable giving.

**JT Finnell** is the director of the Clinical Informatics Fellowship at Regenstrief Institute and is a Research Scientist focused upon operational research with the emergency department. Dr. Finnell was Director of graduate program for Indiana University School of Informatics and computing. Dr. Finnell has extensive experience with curricular development, and has facilitated the involvement of high school students at the Institute. The high school outreach effort at Regenstrief is on an individual student scale. He currently co-directs the AMIA High School Scholars Program.

**Kim Unertl** from the Department of Biomedical Informatics (DBMI) at Vanderbilt University is a researcher focused on the interaction between clinical workflow and health information technology. In addition to her research program, Dr. Unertl works with her department’s high school outreach program and co-directs the AMIA High School Scholars Program. Vanderbilt DBMI currently partners with a local Nashville public high school, Martin Luther King Jr. Academic Magnet High School. This partnership involves recruiting 1-3 students from MLK High School for research internships lasting 8 weeks in the summer, funded through a R25 grant from the National Library of Medicine. Our partnership with the high school allows us to have a direct connection to the students, and we work with administrators and teachers in the intern selection process. Students at MLK High School have the opportunity to take an AP Computer Science course as part of their high school curriculum, which assists us with identifying potential interns. Dr. Unertl has mentored two high school students as summer research interns, both of whom have presented at an AMIA conference.
Neil Sarkar is director of the Center for Biomedical Informatics at Brown University with broad interests in biomedical informatics research and education. He is also the immediate past chair of the AMIA Education Committee, where he first identified the need for AMIA to develop a comprehensive pipeline for a biomedical informatics workforce, akin to how other STEM-C disciplines have approached similar challenges. His research focuses on the development of methodologies for linking biomedical data across heterogeneous sources, leveraging work in ontologies, information retrieval, and natural language processing, with particular application to the study of complex disease phenotypes (e.g., pre-term birth) as well as identification of putative new medical knowledge (e.g., medicinal plants). He is also a co-director of the AMIA High School Scholars Program. He has worked with high school students in the context of biological science, and is a leader for a Brown pre-college summer course aimed at teaching high school students fundamentals of biomedical informatics and data science. Dr. Sarkar will moderate the session, soliciting questions from the audience and guiding the panel discussion.

Discussion of anticipated audience

The intended audience for the panel includes a wide range of AMIA attendees, including researchers interested in educational and mentoring experiences, graduate students seeking mentoring opportunities, individuals involved in directing or managing biomedical informatics educational programs, individuals working with funding agencies and organizations focused on education, and, broadly speaking, any AMIA attendees interested in future directions for the field. The interactive format of the panel makes the panel of interest both to those already engaged in outreach to the high school level and those interested in getting started with outreach efforts. The panel will also welcome conversation around the foundational questions of whether this type and level of outreach is appropriate and important for the field. Based on conversations we have had over the last several years with individuals from across the United States and around the world, the type of cross-organizational information sharing about educational outreach the panel plans to engage in would be of great interest to AMIA symposium attendees and could provide necessary encouragement to expand high school outreach in Biomedical Informatics.

Expected discussion

Although a wide range of topics could be discussed during this panel, we anticipate topics falling into two main categories: foundational and pragmatic. Foundational topics involve fundamental questions around why educational outreach to high school students is important to the biomedical informatics and what kinds of roles biomedical informatics can play in high school educational experiences. These foundational topics could also include discussions around the role of AMIA in supporting or promoting high school outreach activities and about what kinds of funding mechanisms can support these activities. While foundational topics focus more on the why of educational outreach, the pragmatic topics that we anticipate being discussed focus on the what and the how. For example, we plan to discuss the many different approaches that educational programs in biomedical informatics have taken towards designing educational outreach opportunities for high school students, the positive and negative aspects of these different approaches, and the strategies that have worked well for integrating high school students into research opportunities. Over the last several years, we have learned that there are several challenges in engaging high school students in research, including difficulties related to communication styles and interaction expectations. We plan to discuss these challenges and how we have dealt with them as part of the panel discussion. Finally, we anticipate discussing how programs can get started with involving high school students in research experiences and identifying routes to provide meaningful experiences for high school students and their research mentors.

Statement of participation: All panelists have agreed to take part in this panel.

References

Patient Generated Data: the Missing Link in Patient-Centered Care?

Noémie Elhadad, PhD¹, Lena Mamykina, PhD¹, XinXin Zhu, MD, PhD², Eileen Koski, MPhil²

¹Columbia University, New York, NY; ²IBM Watson Research Center, Yorktown Heights, NY

Abstract

In this interactive panel, we aim to engage the audience in an in-depth discussion on the topic of patient-generated data, as captured through smart phones and wearable devices. As patient-generated data receive increased attention from the patient, clinical, and research communities alike, there are numerous synergies and avenues for informatics research. We expect a lively discussion around existing use cases and a diverse set of axes (technological, analytics, human-computer interaction, varied stakeholders).

Introduction

In recent years, patient-generated data have received increasing attention from both clinical and research communities. In the context of health and healthcare, they are commonly defined as data generated outside of the clinical settings that can have an impact on an individual’s health, improve outcomes, and facilitate patient-provider communication¹. From the individuals’ perspective, self-monitoring has long been established as an important component of self-management for many chronic conditions, including asthma, hypertension, and diabetes, among others²,³. Novel technologies provide an unprecedented opportunity to capture and monitor data related to health and wellness. Individuals around the world already use mobile and wearable devices to track their diets, physical activity, sleep, and stress levels. From the healthcare perspective, patient-generated data is often viewed as having a potential to improve treatment and long-term post-treatment survivorship⁴. Perhaps as a result, the Institute of Medicine (IOM) recommends that providers are required to collect standardized social and behavioral data within their Electronic Health Record (EHR)⁵. Finally, under the research agenda of precision medicine and advancing medical scientific knowledge, patient-generated data hold the promise to act as the missing piece in providing a full picture of a patient and capturing data that would not have been possible to acquire in traditional observational settings⁶.

However, together with this general enthusiasm, there remain multiple concerns about patient-generated data and their ability to impact healthcare and self-management or advance medical knowledge. From the perspective of an individual, previous studies raised concerns regarding individuals’ ability to interpret and understand these data and incorporate them into their actions and choices⁷,⁸⁹. From the healthcare perspective, the typical concerns about patient-generated data include those about quality, reliability, and completeness. Moreover, lack of common standards for collection of patient-generated data presents challenges for its integration within EHR, and can potentially lead to duplication of data, if it is collected via multiple channels⁵. Finally, there exist multiple barriers for integrating such data within existing clinical work practices, and temporal and effort-related constraints¹⁰. From the perspective of informatics research, many questions remain as to the availability of computational methods for analyzing patient-generated data.

In this panel, we will present several different perspectives (technological, analytics, human-computer interaction, and the different stakeholders including patients, providers, and biomedical researchers) on patient-generated data and discuss opportunities and challenges related to their wide adoption. The intended audience for this panel consists of informaticists who are either considering including patient-generated data in their ecosystem, building tools for collecting patient-generated data, or platforms to leverage and make sense of patient-generated data.

Aims and Expected Discussion

For this interactive panel, we aim to engage the audience in an in-depth discussion about emerging solutions and approaches to gathering patient generated data for various goals (disease management, monitoring and prevention of disease progression, advancing understanding of disease) and to feeding it back to various stakeholders (patients themselves, healthcare providers, and biomedical researchers). The goal of the panel is to identify common challenges and opportunities across informatics research and practice around patient-generated data. As such, we expect the discussion to evolve around the following axes.
From a technological standpoint, what are the challenges to building tools and solutions for gathering patient-generated data? In particular, many sensors are available to gather a wide range of variables from patients, but most sensors operate on proprietary platforms. Within and outside the informatics community, what are the initiatives to establish standards and models for different types of sensed data streams?

From an analytics standpoint, what is the state of the art for making sense of patient-generated data and turning streams of observations into actionable knowledge? There are several tasks involved towards this goal ranging from low-level tasks like linking often sparse, temporal signals to outcomes to high-level ones, like identifying meaningful sub-groups of patients with respect to outcomes. The characteristics of patient-generated data, including heterogeneous variables, temporality at different time resolution across variables, missing data, accuracy and usefulness of self-reported vs. sensed data (each with their own biases) raise a need for appropriate analytics techniques.

From a human-computer interaction, what are the mechanisms to engage patients and sustain tracking? As a counterpart to engagement, what are the mechanisms to feed actionable knowledge back to patients, for instance to help them learn from the collected data and use this new knowledge to improve individuals’ health?

From the points of the different stakeholders (patients, providers, scientists), what do they want out of patient-generated data and how do they envision it will help striving learning health systems? What are they worried about? As a trigger for discussion, we will explore some use cases on which the panel participants actively work: leveraging a mobile health application to phenotype diseases that are not well understood, such as endometriosis; leveraging smart tools to help diabetes patients learn from their own glucose monitoring streams and nutritional data and manage their conditions, as well as building tools to translate patient-generated data into better clinical insights and actions (identify silent ischemia using sensors, or continuous blood pressure monitors to avoid white coat effect). We will discuss potential pitfalls for each stakeholder. For instance, for providers, information overload needs to be addressed when adding these data to the already complex set of data points available to them. For researchers learning from these data, how to identify the biases of patient-generated data and account for them (whether noisy data from sensors or self-reported data with inherent uncertainty associated to them). For patients, factors such as health literacy, behaviors, and attitudes all impact whether and how they will collect and make sense of their own data and the data of their peers.

Finally, we will expect a discussion from the audience and the panel participants about ways to foster the informatics community interested in patient-generated data and what type of community activities might help propel forward research in this area.

Contributions of the Speakers

Noémie Elhadad, Ph.D. together with her research group develops techniques that aim to support clinicians, patients, and health researchers in their information workflow by automatically extracting and making accessible information from unstructured, large clinical datasets (e.g., the electronic patient record) and patient platforms (e.g., online health communities). She designs novel computational approaches that infer models of health phenomena and translate the learned models into actionable knowledge and robust systems within the healthcare ecosystem. She will discuss the role of data science for making sense of patient-generated data as well as their use for advancing precision medicine and scientific medical knowledge.

Lena Mamykina, Ph.D. has over 10 years of experience with novel technologies for self-monitoring and self-management of chronic diseases. She is a PI on a NIDDK-funded project to design and evaluate a mobile application for facilitating problem-solving in diabetes self-management, that specifically targets individuals from underserved communities and ethnic and linguistic minorities. Dr. Mamykina will discuss opportunities to use patient-generated data to inform and guide individuals’ health self-management.

XinXin Zhu, M.D., Ph.D. is a physician and healthcare informatics professional with more than a decade of experience in wellness management, health informatics research, advisory and consulting, as well as an established track record of developing creative solutions to complex business challenges. Her current research focuses on human-centric health promotion, behavior modification, and patient-provider engagement through telehealth. Dr. Zhu will discuss opportunities for leveraging patient-generated data from the providers’ perspective.

Eileen Koski, M.Phil. is a Program Director for Health Data & Insights at IBM Thomas. J. Watson Research Center. Prior to joining IBM she was with Columbia University (1978-1995), Quest Diagnostics (1995-2009), Medco/Express Scripts (2009-2013) and Northwell Health (2013-2015). Ms. Koski has worked in medical
informatics throughout her career and has been responsible for a broad array of data-related initiatives spanning academic and commercial settings, including population health analytics; data mining; health data visualization and graphical laboratory report design; biosurveillance and pandemic preparedness; clinical trials; enterprise data governance programs; and representing the interests of both the laboratory industry and academic medical centers in numerous government initiatives, task forces, committees and panels as a recognized subject matter expert. Ms. Koski is currently Chair-Elect of the AMIA Knowledge Discovery and Data Mining Working Group. Ms. Koski will be the moderator for this panel.

**Statement from Panel Participants**

All participants have agreed to take part on the panel.

**References**

Women in Informatics Leadership Forum

Rebecca Jacobson, MD, MS 1, Suzanne Bakken, RN, PhD 2, Wendy Chapman, PhD 3, Valerie Florance, PhD 4, and Jessica D. Tenenbaum, PhD 5

1 University of Pittsburgh, Pittsburgh, PA; 2 Columbia University, New York, NY; 3 University of Utah, Salt Lake City, UT, 4 National Library of Medicine, Bethesda, MD; 5 Duke University, Durham, NC.

Abstract Biomedical Informatics is a diverse field encompassing many different areas of study and providing many different potential career paths. Our field has benefited greatly by the intersection of different disciplines which enhance the diversity of Biomedical Informatics as well as the demographics of its contributors. While women have always represented a significant fraction of the discipline, there have been far fewer women in leadership positions in biomedical informatics, particularly in traditional academic roles and professional organizations. Anecdotally, we observe that there are an increasing number of women entering training programs and early academic positions in biomedical informatics. What can we do now to increase the potential for a larger cohort of women leaders in Biomedical Informatics within the next decade? This AMIA Panel will consist of two parts intended to form a single cohesive experience for participants, although participants may choose to participate in either one or both of the related events.

Panel Discussion. The panel discussion will provide brief five to ten minute introductions to the discussion topics by each of the five speakers, followed by discussion of other participants, and then discussion by audience members. Specific topics will include (1) the spectrum of career trajectories and leadership opportunities within them (2) challenges and opportunities for women leaders, (3) mentoring and networking as important success factors and (4) leading within a female dominated field (e.g., nursing informatics) - what’s the same, what’s different, and what can we learn?

Career Guidance Workshop. On the evening following this Interactive Panel, we will hold a 2 hour workshop targeted towards early-career and mid-career women in informatics. The two hour block will be divided into four one-to-one career guidance consultation sessions. Each session will last 30 minutes, and participants may request or be asked to participate in multiple sessions during the workshop. Consultants will be senior and mid-career women with widely ranging titles, roles and interests recruited in advance of the conference. Approximately one month before the meeting, participants wishing to have a career guidance session will submit their Curriculum Vitae or Resume to us along with a ranked list of consultant choices. All participants will be notified of matches in advance of the conference.

Aim of Discussion. The panel discussion and audience interaction components will focus on increasing awareness about potential leadership roles for women. We will discuss challenges for attaining such roles along with practical guidance for overcoming these challenges. An important aspect of the discussion will center on what women in biomedical informatics can do along with professional organizations such as AMIA to increase the ability of women to successfully advance their careers along their chosen trajectories. The topic is of importance to all biomedical informaticists (both male and female) as we endeavor to further expand the pool of future leaders within AMIA and also within biomedical informatics in general.

Intended Audience: The intended audience for the panel discussion includes all biomedical informaticists (both male and female) who interested in enhancing leadership positions for women within our field. Attendees may be primarily interested in better understanding pathways to leadership with their own organizations, or within professional organizations such as AMIA. They may also be interested in identifying resources and strategies for enhancing recruitment and advancement of women into leadership positions in Biomedical Informatics.

Contributions of Each Speaker

Suzanne Bakken, PhD, RN, FAAN, FACMI, is the Alumni Professor of Nursing and Professor of Biomedical Informatics at Columbia University where she directs the Center for Evidence-based Practice in the Underserved
and the Reducing Health Disparities Through Informatics Pre- and Post-doctoral Training Program and is on the Executive Committee of the Irving Institute for Clinical and Translational Research. She currently serves as President of the American College of Medical Informatics and Associate Editor of the Journal of the American Medical Informatics Association. Dr. Bakken will compare and contrast her lessons from leadership positions in nursing and from interdisciplinary initiatives.

Wendy Chapman, PhD, FACMI, is the chair of the Department of Biomedical Informatics at the University of Utah. She discovered biomedical informatics serendipitously on a journey from Elementary Education to Linguistics to Chinese Literature and fell in love with the field. Since receiving her PhD in 2010, Dr. Chapman was a faculty member at the University of Pittsburgh then UC, San Diego. She is the chair of the AMIA Student Paper Awards Committee, and is an elected member of the AMIA Board of Directors. Dr. Chapman will discuss leadership of academic informatics units

Valerie Florance, PhD, FACMI is an Associate Director of the National Library of Medicine, NIH. There, she directs NLM’s extramural grants program. She also serves as program director for NLM’s highly-regarded university-based training programs in biomedical informatics. Dr. Florance is a member of the Executive Committee for the NIH Big Data to Knowledge (BD2K) initiative. Before coming to NLM in February 2001, she spent 3 years leading a visioning project at the Association of American Medical Colleges (AAMC) to help the association’s members understand the power of computers and networks for managing health information. Before that, she held faculty and administrative positions at three U.S. academic medical centers. She has graduate degrees in medical anthropology and biomedical information sciences, and was elected to the American College of Medical Informatics in 2005. Dr. Florance will focus mainly on informatics leadership options in non-academic settings.

Rebecca Crowley Jacobson, MD, MS, FACMI is Professor of Biomedical Informatics at the University of Pittsburgh. She is the Director of the NLM-funded Pittsburgh Biomedical Informatics Graduate Training Program, and Chief Information Officer for the University of Pittsburgh Institute for Personalized Medicine. She is also the incoming Chair of the Biodata Management and Analysis (BDMA) NIH Study Section. Dr. Jacobson will introduce the panel discussion. She will also discuss the critical role of networking and mentoring.

Jessica Tenenbaum, PhD is an Assistant Professor in Duke's Department of Biostatistics and Bioinformatics, Division of Translational Informatics. She plays a leadership role in the American Medical Informatics Association, serving as Chair of the Genomics and Translational Bioinformatics Working Group and as an elected member of the Board of Directors. She is an Associate Editor for the Journal of Biomedical Informatics and serves on the advisory panel for Nature Publishing Group's Scientific Data initiative. After earning her bachelor's degree in biology from Harvard, Dr. Tenenbaum worked as a program manager at Microsoft Corporation in Redmond, WA for six years before pursuing a PhD in biomedical informatics at Stanford University. She was the Associate Director for Bioinformatics for the Duke Translational Medicine Institute for 7 years before taking on a faculty role at Duke in 2015. Dr Tenenbaum will address the importance of, and best practices for, networking and what she views as the most valuable takeaways from Sheryl Sandberg’s well known book Lean In.
Global e-Health: The Latin American Perspectives

Heimar F. Marin, PhD, FACMI1, David Novillo, PhD2, Alexandre Barbosa, PhD3, Fernan Gonzales B. De Quirós, PhD, FACMI4

1Universidade Federal de São Paulo, São Paulo, Brazil; 2Pan American Health Organization, Washington DC, USA; 3Regional Center for Studies on the Development of the Information Society – Cetic.br, São Paulo, Brazil; 4Hospital Italiano de Buenos Aires, Argentina.

Abstract

Health systems around the world face considerable challenges in providing health care services. Governments of the Latin America region have also realized the importance of e-Health and the need to create adequate policies and national strategies that could facilitate international and regional cooperation in this area. This panel will discuss the vast possibilities of ICT in health care focusing on research and initiatives developed by countries in Latin America to influence policy makers for fostering ICT use in health care to improve health conditions for the population.

Introduction

Advances in science and technology in global healthcare demand innovative solutions to address complex problems that are common across countries such as the aging of the population, chronic diseases, high costs of care delivery, economic constraints to assure continuity, quality, and cost benefits of treatments, including care for emerging and re-emerging diseases. To face this scenario, countries are seeking to improve their health systems, enhancing the efficiency and operative capacity of the national health programs, decreasing mortality and morbidity and improving the quality of life through informatics based systems. Governments of the Latin America region have also realized the importance of e-Health and the need to create adequate policies and national strategies that could facilitate international and regional cooperation in this area. Thus, considering the challenges that Latin America and Caribbean countries are facing, information and communication technology (ICT) is a key instrument that potentially can support effective, equitable and efficient way to improve access to healthcare services. ICT can increase the availability of healthcare resources by optimizing processes, shortening distances, reducing patient transportation, providing timely care and cutting costs for families and the health system.

Health care systems models are affected by several factors, including aging of population and the consequent change in the way that health services are requested, changing on epidemiological profile of populations (decrease of infectious diseases with higher prevalence on non-communicable diseases, resurgence of infectious diseases of the old agenda such as tuberculosis, dengue, cholera, and the emergence of new infectious diseases such as AIDS, Ebola and hantaviruses), significant alteration on the way knowledge is diffused that generate important modifications in the willingness of people related to consume health services. It can also mention the health professional’s models of working by division and tasks that deconstruct the necessary inter and trans-disciplinarily in health and turns the patient into a victim of multiple professionals. The technological revolution (not ICT) that offers new alternatives without necessarily abandoning the previous one, causing extra costs1.

Among all factors, aging of population can be highlighted as an example of how each of these factors affect the whole healthcare system. Due to the progressive aging, people’s greatest expenditure on health occurs during the final decade of their lives. For this reason, health systems allocate a significant part of their resources to the provision of curative and palliative services2. Within this scenario, countries in Latin America and the Caribbean are expected to increase their expenditure on health between 3% and 9% of GDP by 2040. In 2014, IDB has warned that public healthcare spending in Latin America is not consistent with the current demand and as a consequence, governments would face financial burdens over the decade. The conclusion was obtained from analyses of healthcare benefit plans in seven countries – Argentina, Chile, Colombia, Honduras, Mexico, Peru and Uruguay3. On the other hand, spending more
does not guarantee quality and equity in healthcare access and better life conditions. Strategies for development of e-health, policies, and evaluation measures play a key role in promoting adoption of e-health technologies and cost-effective development and implementation. Efficient use of ICT in health care can be translated in terms of clinical, organizational ad financial benefits. In general, some are more tangible to measure. Most recent studies show that financial benefits are not immediate postponed after the initial deployment. Systems that were implemented during the last 40 years, now are able to demonstrate efficiency and maturity to consider their rapid generalization.

Considering those factors and the need to evaluate e-health scenarios in the region, this panel will discuss the vast possibilities of deployment, focusing on research and initiatives developed by countries in Latin America to influence policy makers for fostering ICT use increasing the sustainability of projects to advance management on the national health systems.

**Objectives**

These panels intend to discuss political and societal aspects of health information technology in the Latin America region. The presentations will briefly discuss experiences based on regional countries and PAHO/WHO recommendations but will explicitly draw generalizable conclusions intended to provide insights and information for those who attend the session. The goal is to equip attendees with ideas and recommendations for how to improve eHealth strategies in their own environments.

**Panel Description**

Dr Marin will introduce the theme addressing how the strategies for development of e-health, policies, and evaluation measures play a key role in promoting adoption of e-health technologies and cost-effective development and implementation. Efficient use of ICT in health care can be translated in terms of clinical, organizational ad financial benefits. In general, some are more tangible to measure. Examples will be provided by most recent studies that show financial benefits are not immediate postponed after the initial deployment and by systems that were implemented during the last years that now are able to demonstrate efficiency and maturity to consider their rapid generalization.

Dr Novillo will address the main actions taken by PAHO, among which are research, capacity building, creation of technical guidelines and building knowledge networks. A lack of political commitment is one of the main obstacles to implementing eHealth-based systems and services. Political commitment is key to realizing its full potential, through sustainable, appropriate, and integrated implementation of eHealth initiatives. In the region of the Americas, almost all the countries have plans to use ICTs in the health sector. This does not necessarily mean that all have developed national eHealth strategies or plans. PAHO has seen that political commitment and the adoption of measures on eHealth have increased steadily, encouraging resource mobilization to adopt and develop digital services that allow better access, expand coverage and increase financial efficiency of the health care systems.

Dr Barbosa will describe how since 2012 Cetic.br has worked with an OECD Expert Group responsible for developing a model of the survey questionnaire and a set of indicators to approach ICT adoption, availability and use in the health sector. Based on these efforts and on the experience accumulated in the development of first survey, Cetic.br has also participated in the ICT Work Group from the Statistical Conference of the Americas (SCA) of the Economic Commission for Latin America and the Caribbean (ECLAC) in the effort to establish a common module for measuring the adoption of ICT in the Latin American healthcare sector. This model was approved in 2014 and contains methodological directives and a model questionnaire for disseminating this methodology, which can help guarantee the future production of comparable indicators in the region. After Brazil, Uruguay was the second country in Latin America and the Caribbean to adopt the methodological framework created by OECD.

Dr Fernan Quiros will address the 20 year experience of developing and implementing a Health Information Systems in a complex scenario of a not for profit academic medical center such as Hospital Italiano de Buenos Aires. This experience has been extended to other institutions across Argentina, Uruguay and Chile. The lessons learned from this experience could be useful to others developing countries, since the challenges that the healthcare system at national level are very similar all throughout this countries that have not completely embraced the concept of the benefits that ICT can provide. He will also describe the common pitfalls that most institutions both public and private need to address in order to integrate both the clinical and administrative layers for a successful implementation. In other aspects, the involvement of patients as main actors of a HIS that would benefit their well-being and the relationship with the healthcare team, can also help to establish the base for a Learning Health System that can be extended to a local, national and regional level.

**Panel Structure and Participants**

265
The panelists will offer prepared remarks to introduce the theme and share their experiences across countries. Dr. Marin, as the moderator will introduce the topic and assure that time allocations are respected by all participants.

Panelists

Joining Dr. Marin are three panelists with experience and expertise in the area of e-health, policies and biomedical informatics:

Heimar de Fatima Marin, RN, MS, PhD, FACMI is Professor and Director of the Graduate Program in Management and Health Informatics, Universidade Federal de São Paulo. She is also Editor-in-Chief of the International Journal of Medical Informatics and the Scientific Coordinator of the ICT Health Survey in Brazil.

David Novillo, MLIS, PhD is the coordinator of the eHealth Program of the Pan American Health Organization, Regional Office of the World Health Organization for the Americas (PAHO / WHO) in Washington, D.C. He also leads the WHO eHealth portfolio in Latin America and the Caribbean (LAC), providing technical, managerial and policy advice in areas such as knowledge management, telemedicine, mHealth, electronic health records, standardization and interoperability of health information systems, digital literacy, and social media for public health.

Alexandre Barbosa, MS, PhD is the head of the Center for Studies in Information and Communications Technologies (CETIC.br), an Unesco Category II Center for the development of information society based in Sao Paulo, Brazil. Responsible for several nationwide ICT Survey projects in Brazil aimed at the production of ICT-related statistics and indicators.

Fernan Quiros, MD, FACMI is the Vice Director of Strategic Planning in the Hospital Italiano de Buenos Aires (HIBA) Argentina where he has developed research and educational programs in Medical Informatics, Disease Management, Internal Medicine and Physiology. He is the head of the Area of Research at the Internal Medicine Division. He has created a 4-year Medical Informatics Residency Program that started in 2001 at HIBA. More than 20 fellows have been trained in these programs under the supervision of the Department of Research and Education at HIBA. Dr. Quiros' research focuses on the development of electronic medical records and disease management programs.

Expected Attendees

Although the panel is intended for those who are involved with research, governance and policies in the use of e-health, the topic is of interest of the health informatics communities since we will discuss current aspects on organizations policies guidance that will impact locally on decision making process and policy makers.

References

2 Andrés Fernandez & Enrique Oviedo (eds). e-Health in Latin America and the Caribbean: progress and challenges. ECLAC, Santiago, Chile, 2011.
3 ECLAC, Economic Commission for Latin America and the Caribbean, 2010. Population and Health in Latin America and the Caribbean: Outstanding Matters, New Challenges LC/L.3216(CEP.2010/3), Santiago, Chile.

Statement of the Panel Organizer

All participants have agreed to take part on the panel through its conclusion. Heimar Marin, March 6th, 2016.

Address for Correspondence:
Heimar de Fatima Marin
Rua Dep Bady Bassit 440 – CEP 05517-050 São Paulo – SP – Brasil
Panel Sponsored by Evaluation and People & Organizational Issues Working Groups

A Forum on Qualitative Research in Biomedical Informatics: Controversies, Challenges, and Opportunities

Laurie L. Novak, PhD¹ (moderator); Rupa Valdez, PhD²; Tiffany Veinot, PhD³; Jan Talmon, PhD, FACMI⁴; Nancy M. Lorenzi, PhD, FACMI¹

¹Vanderbilt University, Nashville, TN; ²University of Virginia, Charlottesville, VA; ³University of Michigan, Ann Arbor, MI, ⁴HI-way, Sint Odilienberg, The Netherlands

Abstract

Qualitative research has a long history in biomedical informatics. It has often been used in investigations regarding the social influences on emerging technologies, the impact of technology on individuals and groups, and in improving design and implementation of informatics tools. However, researchers using qualitative approaches may face controversies and challenges in obtaining funding, interfacing with researchers using different methods, and in communicating their work. This panel will engage the audience in addressing critical issues facing qualitative researchers and discussing emerging opportunities that impact the direction of the field. Panelists will briefly discuss their experiences in the following areas: Dr. Lorenzi will discuss organizational qualitative research; Dr. Valdez will discuss home- and community-based qualitative research; Dr. Veinot will discuss funding strategies and Dr. Talmon will discuss issues and opportunities for publication of research findings. The subsequent interactive discussion with the audience will build on the panelist comments, creating a forum to explore a variety of challenges and opportunities.

Intended Audience

The intended audience for this panel includes all AMIA attendees who are interested in the state and direction of qualitative research within the informatics community, those who are consumers of qualitative research such as technology designers and implementers, and individuals who seek to collaborate with qualitative researchers.

Introduction of the Topic

Qualitative research in biomedical informatics has existed in tension with both the assumptions of technology designers [1] and the positivist foundations of biomedical science [2,3]. Elsewhere these tensions have recently manifested in a vigorous debate about the utility of qualitative research as assessed by the publication frequency of qualitative papers in a leading medical journal [4,5]. Qualitative researchers reckon with these issues in their everyday work as they describe the details and benefits of qualitative approaches to funding agencies, coach colleagues on using the methods, and develop new, innovative ways to improve efficiency and value of the methods. Meanwhile, the field of medical informatics is becoming heavily influenced by methodological trends in data science, an increasing focus on patient-centeredness and precision medicine, and the emergence of new resources for research ranging from genomic data to social media. Given this dynamic landscape, qualitative researchers have requested more time at AMIA to share experiences and discuss the place of qualitative approaches within our field. This panel will be a forum for making sense of controversies, identifying challenges and ways to address them, and sharing information about opportunities on the horizon.

Aim of the Discussion

To discuss controversies, challenges, and opportunities associated with qualitative research in biomedical informatics, and to produce insights for future directions based on the expert panel and audience input.

Panelist Contributions
Nancy M. Lorenzi, PhD, FACMI, will discuss qualitative research before, during, and after the implementation of information technology in health care settings. Understanding the people-process-organizational impacts of technology creates the potential to influence or change implementation strategies for greater success. Methods that build on prior informatics research in organizations [6] present opportunities for rapid assessment and “diagnosis” of organizational and technological issues. However, while these studies are critical to better understanding the impact of technology on organizations they can also be controversial [7]. The qualitative researchers are often employees in the organization they are studying and want to maintain good relationships with colleagues who may be participants in interviews, focus groups, and observations. A continued positive relationship with the vendor might be critical for future collaborative work. An in-house organizational researcher faces many challenges such as making time available to implement the research, coordinating the research study with the technology implementation, working with health system operational units while maintaining the rigor of research methods, and reporting both positive and negative findings to the organization and to the greater public.

Rupa Valdez, PhD, will discuss qualitative research to guide the design of consumer health informatics applications intended for use in home and community settings. Achieving patient-centered solutions requires deeply engaging all individuals involved in a patient’s self-management process, including healthcare professionals, informal caregivers, and care partners. Common challenges of this deep level of engagement through qualitative approaches include gaining access to the population, developing the trust and relationship necessary to allow for prolonged interaction during an iterative design process, and the time required by both the researcher and participant to participate in data collection and analysis. New technological developments including social media and automated text analysis present opportunities for mitigating these challenges. Research approaches drawn from fields such as global development, including community partners as co-PIs, also have the potential to promote new ways of engaging in community-based informatics research.

Tiffany Veinot, PhD, will discuss the challenges of obtaining funding for qualitative research in health informatics. Researchers working primarily from a qualitative paradigm may struggle to obtain needed funding to support their research. There are few funding opportunities that explicitly solicit research using qualitative methods, and review panels may struggle to see the value of proposals for projects with a large qualitative focus. This presentation will focus on strategies for conceptualizing and proposing qualitative research to funding agencies. Such strategies include careful selection of funding opportunities, the formation of multidisciplinary teams, clear delineation of the contribution of qualitative data collection/analysis in a larger project, methodological innovation, and design of research that moves beyond descriptive goals.

Jan Talmon, PhD, FACMI, will discuss publication of qualitative research in biomedical informatics. Much of the research published in leading biomedical informatics journals is still about algorithms and methods, which may or may not become part of routine clinical practice. Some of these journals are more interested in how these tools work in practice. There are, however, challenges in this approach. Although quantitative approaches may help to understand the extent of the impact of ICT in health care; such studies may have less explanatory power regarding why such tools do or don’t work nor on the unintended consequences. Qualitative studies are designed to shed light on those issues. The major challenge to get such research published is that it often reflects the local situation and results may not necessarily be of more general interest. For this reason, emphasizing broader theoretical implications is a critical aspect of communicating qualitative research results.

Expected Discussion

After the introductory presentations, the panel and audience will engage in a discussion of various topics, facilitated by the moderator. Topics relevant to this panel are:

- What are the major challenges to obtaining funding for qualitative research?
- How can research proposals be written to emphasize the value of qualitative and mixed methods research for a given problem?
- What are the major challenges to publishing qualitative research?
- How can researchers clearly present the transferability of their qualitative research findings to others?
- How can we clearly articulate the value of the qualitative research in mixed methods studies?
• How can qualitative methods integrate with data science methods?
• How can approaches to making qualitative analysis more “efficient” be integrated with more traditional approaches to qualitative analysis?
• How can qualitative field studies be “scaled up” for large implementation studies?
• What balance should be achieved between engaging deeply and in person with a local population and engaging broadly and virtually with a more geographically dispersed population?
• What are the implications of larger debates about publishing qualitative health sciences research for the biomedical informatics community?

Timeliness of the Topic

This panel is timely because rapid changes in the field of biomedical informatics have created new opportunities for funding and collaboration, and qualitative research is considered to be a key component of major programs such as the Precision Medicine Initiative. Widespread implementations of electronic health records in many organizations have also created awareness of the need for qualitative studies that incorporate organizational context into data gathering and analysis. Similarly, the proliferation of consumer facing health IT solutions that are often abandoned after initial use has created awareness of the need for studies that deeply engage patients and others involved in their care at home and in the community. Yet, despite these increased demands for qualitative research, barriers to funding, execution, and publication remain. The sharing of experiences stimulated by this panel will serve as a foundation for methodological advancements, collaboration, and advocacy for rigorous qualitative research within biomedical informatics.

All panel participants have agreed to take part in this panel.

References

5 Loder E, Groves T, Schroter S, et al. Qualitative research and The BMJ. BMJ 2016;352. doi:10.1136/bmj.i641
7 Lorenzi NM, Riley RT. Organizational issues = change. Int J Med Inf 2003;69:197–203.
Population Health Informatics: Connecting Consumers to Care

Harm J. Scherpbie, MD MS\textsuperscript{1}, Vincent A. Emanuele II, PhD\textsuperscript{2}, Jane Sarasohn-Kahn, MA(Econ.), MHSA\textsuperscript{3}, Mark Scrimshire\textsuperscript{4}

\textsuperscript{1}Jefferson College of Population Health, Philadelphia, PA; \textsuperscript{2}Wellcentive, Atlanta, GA; \textsuperscript{3}THINK-Health, Valley Forge, PA; \textsuperscript{4}Centers for Medicare & Medicaid Services, Baltimore, MD and Medyear, New York, NY

Abstract

This Interactive Panel Discussion explores the challenges to use Informatics tools and methods to connect patients and consumers to their care givers and make them active participants in their care. In this panel we bring together perspectives from providers, consumers, data science, and technology firms, to engage with the audience participants to discuss technology approaches for connecting consumers and patients to their care: How far we have come? What are the challenges ahead? What are the obstacles that we need to overcome? What will it take to bring the consumer role in health care to the same level as the consumer in retail, entertainment, and other industry sectors? And most importantly, what will that do to the quality and efficiency of our care?

Introduction

This Interactive Panel Discussion explores the challenges for using Informatics tools and methods to connect patients and consumers to their care givers, and make them active participants in their care. One of the pillars of the Meaningful Use program was “Patient and Family Engagement”, and this created a surge in Patient Portals, View/Download/Transmit capabilities, Secure Messaging options between patients and providers, and targets for providers to achieve patient participation in these initiatives. Patient and Family Engagement is also one of the major components of any Population Health initiative: most Population Health Management Systems include a module or method for Patient Engagement. The result of these patient engagement initiatives are mixed, and while many patients use portals, and communicate with their providers, the majority do not. There likely are multiple causes, or reasons, for the lagging adoption – technical, psychological, organizational, cultural, economic, and more. This panel discussion explores the challenges and obstacles for using information systems and tools for population health – from the perspective of a provider organization, the consumer, the technology firm, and the population health data scientist.

Panel Members and Perspectives

Harm Scherpbie MD MS (Moderator). From his perspective as prior CMIO for a large health system, Dr. Scherpbie focuses on the provider side of the customer / care relationship. EMR systems are now widespread in physician practices, hospitals, and health systems. The next pressure on providers is the transition from Fee for Service-based care to Fee for Value – the transition to Population Health. Harm will discuss how providers plan to use information systems to manage care and reimbursement under the new rules, and how they will use information systems to communicate with and care for patients and consumers.

Vincent Emanuele PhD. Dr. Emanuele is a data scientist at Wellcentive, a Population Health Management firm. He will explore what population health systems, from a data perspective, tell us about patients and consumers, and how we can use data and analytics to do a better job in identifying consumers and engaging patients. What do claims data and EMR-derived clinical data tell us about populations, and about the providers who care for them? How do socio-economic data help characterize segments of the population, and how does this affect the patient engagement strategies? Data scientist look at this from the highest level, and Vince will bring this perspective to the conversation.

Jane Sarasohn-Kahn, MA, MHSA. Mrs. Sarasohn-Kahn is a Health Economist and Advisor, a frequent blogger and social media participant, with a large number of followers, and from this platform she monitors trends in consumer approaches to health care. The spectrum of consumer engagement and participation goes from the highly connected and engaged, to the hard to reach, disengaged patients who may or may not need care. What determines where a consumer is on this spectrum? – and how can we move people toward engagement? Jane’s perspective includes behavioral, societal, cultural, but also technical and social-media related aspects of consumer behavior in healthcare.

Mark Scrimshire. Mr. Scrimshire comes from the perspective of the technology industry as CTO of Medyear, a start-up Population Health Record firm. As Entrepreneur-in-Residence at CMS, he is familiar with the policy aspects of technology in healthcare. Mr. Scrimshire will discuss how BlueButton+ and FHIR can move us from EMR-tethered
portals (current state) to patient-tethered portals and patient-controlled health records. Mark will discuss his perspective on the technology industry’s ability to deliver on this promise, and whether consumers will be ready to take up this new technology.

Most important for this panel discussion will be the role and participation of the audience. The panelists’ introductions will be short and to the point, and the focus of the discussion will be the questions and comments from the audience. These three factors make this panel discussion successful and timely:

- Everyone in health-care today can relate to this topic – whether you work for a provider organization, research institution, technology firm, or consumer-oriented organization. Everyone has an experience and an opinion on this topic, and we look forward to a forum where these topics come to the floor.

- The transition to population-based care, the rapid shift in technology, increasing interoperability in EMRs and the removals of data blocks put us in a time of rapid change. This is promising and threatening at the same time, and we need to consider the challenges and obstacles ahead.

- Panelists representing providers, consumers, technology firms and data scientists can share their perspectives and different points of view.

**Conclusion**

Consumers and patients will play a strong role in their care. Healthcare providers transition to population-based care models and value-based payment models. Both of these trends are reliant on data and information technology. This interactive panel discussion engages participants to discuss the future outlook, and the road ahead to connect consumers to their care.
Interactive panel:
How informatics are being applied in industry: challenge and opportunity

Xia Wang, PhD1, Jorge Caballero, MD2, Zhaohui Cai, MD, PhD3, Elizabeth S. Chapman, MS, CPHIMS4, Shahram Ebadollahi, PhD5, Matvey B. Palchuk, MD, MS6

1 AstraZeneca Pharmaceuticals, Gaithersburg, MD, USA; 2 Distal, Menlo Park, CA, USA; 3 Celegene, Summit, NJ, USA; 4 Department of Veterans Affairs (VA), Jackson, GA, USA; 5 IBM Watson Health, New York, NY, USA; 6 TriNetX, Inc., Cambridge, MA, USA

Abstract: Due to its multi-disciplinary nature, informatics is a field with applications in a variety of settings, among them the secondary use of health care data in industry represents exciting opportunities and challenging career paths for informaticians. These individuals are required to possess innovative thinking and mastery of the cutting-edge technologies, to work on complex, real-world problems in areas such as big data mining, novel treatment development, precision therapies, clinical decision support, mHealth applications and device data analytics, with the ultimate goal of each patient having access to the highest quality and personalized care options. On this panel, representatives from AMIA’s corporate members will share various use cases and discuss how the informatics innovations are being applied in industry to address challenges and make break-throughs. The panel will also share their own career path and how they arrived at their current positions, discuss the important characteristics they are looking for in career seekers and the challenging opportunities residing within the industry. Following individual presentation, the panelists will take questions from the audience on a range of topics related to how informatics is being applied and the role of informaticians in industry.

Learn objectives:

1. Understand the specific applications of informatics innovations in various health-related industries
2. Identify the unique challenges and suitable opportunities in industry to start a career or execute a career transition
3. Obtain the recommendations for developing skills and value-creating innovative thinking that are critical for practicing informatics in industry

Panel description:

The panel will focus on specific use cases of how informatics is being applied in diverse health-related industry settings. The panelists will present the unique challenges regarding secondary use of healthcare data in industry and how informatics innovations are being applied to make break-throughs. During the presentations, the panelists will also talk about their own career paths and how they landed in their current position. They will share their perspectives on exciting
opportunities in industry for informaticians and provide constructive recommendations for audience via interactive discussion.

**Introduction**: Xia Wang, PhD, will introduce the topic, panel and moderate the session.

**Presentations 1**: Creating a data ecosystem that makes health information easily accessible and meaningful for everyone. Jorge Caballero, MD, will discuss why individuals with informatics skills sets that are simultaneously broad and deep are poised for success in rapid-growth environments. He will also share how Distal relies on its team of clinicians, scientists, and engineers to enable data-driven care with a personal touch. To close, he will offer specific suggestions for those looking to apply their informatics training to overcoming the challenge of delivering sophisticated, personalized data analyses to a diverse audience on-demand, and at scale.

**Presentation 2**: Developing a Learning Healthcare System – what’s the role for informaticians in the biopharma industry? Zhaohui “John” Cai, MD, PhD, will discuss on a vision for pharma to leverage big healthcare data for patient benefit by working together with other sectors in the healthcare ecosystem – the biopharmaceutical company of tomorrow is one that is not only able to excel in product discovery, development, and commercialization, but one that is able to accelerate, optimize, and transform its core activities by embracing advancements in data and technology. He will share how the Celgene Medical Informatics team is helping the company to achieve that vision through a combination of partnership strategies, innovative approaches and analytics, as well as leadership and career development specific to informaticians.

**Presentation 3**: Boots on the ground – the battle on the front lines. Elizabeth S. Chapman, MS, CPHIMS, will discuss the interprofessional nature of informatics and the challenges and opportunities at the point of care. Despite the Department of Veterans Affairs’ (VA) long and storied history in the application of information technology in health care, usability, workflow integration, interoperability and mining the vast amount of data to yield actionable insights remains challenging. Ms. Chapman will discuss ongoing efforts in VA to define the informatics workforce and equip staff with the requisite knowledge and skills to implement and support information technology that improves healthcare quality, safety and efficiency.

**Presentation 4**: Watson Health: Enabling Computational Health in the Era of Big Data. Shahram Ebadollahi, PhD, MBA, will provide an overview of IBM Watson Health and then will focus on the innovations leading to establishing a new business unit focused on analytics and health informatics within a large corporation. He will share his views and experience in the role of informaticians, their importance and place in multi-disciplinary teams required for innovating in the area of analytics in healthcare applications.

**Presentation 5**: Accelerate clinical trials design and recruitment – from art to data-driven science. Matvey B. Palchuk, MD, MS will focus on secondary use of clinical data for research. He will describe how TriNetX is relying on applied informatics to enable queries across a globally-distributed federated database of heterogeneous clinical data. Dr. Palchuk will discuss challenges of working with various sources of data, mapping data to standard coding systems, assessing data quality, creating innovative analytics and aiding users in designing, executing and interpreting results of patient cohort identification queries. He will highlight the role of informatics professional in a startup company targeting innovative data-heavy solutions in healthcare and life science markets.
Panel organizer and participants

Xia Wang PhD, Principal Scientist and informatics lead, Global Medicines Development, AstraZeneca Pharmaceuticals, Gaithersburg, Maryland, USA

Jorge Caballero, MD, CEO of Distal, Menlo Park, CA, USA

Zhaohui Cai, MD, PhD, Director, Medical Informatics, Celgene, Summit, NJ, USA;

Elizabeth S. Chapman, MS, CPHIMS, Workforce Development Competency Lead, Veterans Health Administration, Office of Informatics and Information Governance, Department of Veterans Affairs (VA), Jackson, GA, USA

Shahram Ebadollahi, PhD, VP, Innovation & Chief Science Officer, IBM Watson Health, New York, NY, USA

Matvey B. Palchuk, MD, MS, VP, Informatics, TriNetX, Inc., Cambridge, MA, US

Statement of the Panel organizer

All participants have agreed to take part on the panel.

This panel was vetted through the IAC (Industry Advisory Council) and all panel members are also members of the IAC.
Personalized and Precision Medicine At Scale

Lewis James Frey, Ph.D\textsuperscript{1}; Danny Sands, MD, MPH, FACP, FACMI\textsuperscript{2}; Nephi Walton, MD, MS\textsuperscript{3}; Ken Yale, DDS, JD\textsuperscript{4}

\textsuperscript{1}Medical University of South Carolina, Charleston, SC; \textsuperscript{2}Harvard Medical School and Society for Participatory Medicine, Boston, MA; \textsuperscript{3}Washington University, St. Louis, MO; \textsuperscript{4}ActiveHealth Management/Aetna, New York, NY

Abstract

Personalized and precision medicine has been an interesting topic of discussion, with few examples of large-scale, replicable, population-wide programs - until now. In 2014 Aetna completed one of the first health-plan sponsored programs using genetic testing in a large population to enhance diagnosis and target treatment for persons at-risk for metabolic syndrome. The original pilot was based on sound science, however the shortcomings of current knowledge on the use of genetics for diagnosis and treatment are apparent. Not as obvious, however, are the proven benefits to population health. Based on published findings from the pilot and lessons learned, we are refining the original study and expanding to other conditions. Goals of this panel include: discuss the different genetic tests useful for metabolic syndrome and other chronic conditions, determine appropriate pre-screening procedures using big data and novel predictive analytics, understand the probabilistic nature of genetic variants and implications for precision medicine, acknowledge provider, payer, and patient skepticism of the promise of genetic testing, and identify other conditions currently appropriate for personalized medicine studies base on existing genetic knowledge and return on investment. This submission is sponsored by the AMIA Genomics and Translational Bioinformatics Workgroup.

Introduction - Why the Panel is Timely and Controversial

Metabolic syndrome is a serious and growing global health problem. Persons with metabolic syndrome are twice as likely to develop cardiovascular diseases, and five times more likely to develop diabetes mellitus. Thus, the ability to more easily identify persons at-risk for metabolic syndrome, and intervene with appropriate wellness options, is critical. The original pilot was a first-of-its-kind program using a novel predictive analytics approach to identify candidates appropriate for further testing, and genetic testing to intervene in a condition in a large-scale setting. The pilot resulted in significant advancements in identification of persons with metabolic syndrome and intervention to prevent further morbidity and reduce costs.

There are, however, shortcomings to the study, some of which can be addressed going forward, but others are inherent in this nascent field and may not be avoidable. To begin, there is a relatively low probability of the few (three) common genetic variants indicative of metabolic syndrome to actually result in obesity or metabolic syndrome. In addition, effectiveness in widespread population screening for certain chronic conditions has been questioned, as there are usually many genes involved in a disease state, each having a small effect. As a result, it is not always clear whether any particular genetic variant (whether SNP, copy number variant, or insertion/deletion) is harmful, and what specific action should be taken. Some suggest whole genome sequencing is more important to identify persons at-risk for chronic conditions, but that is still expensive – and it is unclear how persons will react or know how to evaluate risks or probability of a condition. Nevertheless, significant benefits were identified for persons who participated in the pilot program, in both improved wellness and reduced costs. Moreover, personalized and precision medicine cannot wait for the perfect program to be developed, you have to start somewhere, so any improvement in health quality, outcome, or cost is beneficial and will move the field forward. The plan is to expand the pilot to more persons and additional conditions. The question is how best to accomplish this task given the current state-of-art in data and genetic science.

Personalized and Precision Medicine Pilot

We designed and implemented a novel predictive model for metabolic syndrome risk to predict subsequent risk of metabolic syndrome at both population and individual levels.\textsuperscript{2} Once identified, persons with elevated risk were invited to participate in a targeted treatment program, which included a limited genetic profile, a traditional psychosocial assessment, and high-intensity coaching in a randomized controlled study. Results of the study included sustained patient engagement, weight loss, improved clinical outcomes, and reduced health care costs. We also learned important lessons in patient behavior and willingness to participate and engage in genetic testing and
programs designed to improve health and wellbeing. Reduced health care costs are critical for the return on investment necessary for personalized and precision medicine to make sense economically and result in widespread adoption. We anticipate similar results for the other chronic conditions under consideration.

**Intended Audience:** Providers, Healthcare Executives, Clinical Geneticists, Health Plan Executives, Data Scientists, Entrepreneurs, anyone considering implementation of a large-scale personalized or precision medicine project with a defined return on investment

**Contribution of Each Speaker**

Dr. Lewis Frey – Moderator, shall give an overall introduction, and mediate between panel participants

Dr. Ken Yale – Proponent, who represents the aspirational perspective of personalized medicine, and shall emphasize the potential for genetic testing to intervene in chronic conditions in large-scale, population health settings, significant advancements in identification of persons with metabolic syndrome, intervention to prevent further morbidity and reduce costs, and potential for other chronic condition diagnosis, early intervention, and treatment.

Dr. Danny Sands – Physician skeptic, who represents perspective of the physician community regarding how genetic information is presented to patients, concerns of patients about how best to use genetic data in their own care, and the lack of clarity about actions to take with the information.

Dr. Nephi Walton – Genetics expert and clinical informaticist, who represents the perspective of clinical genetics in practice, challenges faced by patients, and the need for greater education of patients, providers, and payers about the promise and limits of genetic tests. He shall discuss the promise of precision medicine in light of the limits of our current state of genetic testing and knowledge about common diseases. In addition he will address the inadequacies of our current health IT infrastructure to process and store genomic information and the work that needs to be done to remedy these problems.

**Conclusion** - Personalized and precision medicine hold great promise, but also peril. Early results of large-scale implementations in real-world population health environments are promising. But there is significant misunderstanding among providers and the public about the benefits and drawbacks of using genetics to diagnose and treat patients. Further education and discussion is needed to inform the appropriate use of precision medicine at scale.

**References**


Expanding access to high-quality plain-language patient education information through context-specific hyperlinks

Jessica S. Ancker, MPH, PhD,1 Elizabeth Mauer, MS,1 Diane Hauser, MPA,2 Neil Calman, MD2

1. Department of Healthcare Policy & Research, Weill Cornell Medical College, New York, NY
2. Institute for Family Health and Department of Family Medicine and Community Health, Icahn School of Medicine at Mount Sinai, New York, NY

ABSTRACT
Medical records, which are increasingly directly accessible to patients, contain highly technical terms unfamiliar to many patients. A federally qualified health center (FQHC) sought to help patients interpret their records by embedding context-specific hyperlinks to plain-language patient education materials in its portal. We assessed the impact of this innovation through a 3-year retrospective cohort study. A total of 12,877 (10% of all patients) had used the MPC links. Black patients, Latino patients comfortable using English, and patients covered by Medicaid were more likely to use the informational hyperlinks than other patients. The positive association with black race and Latino ethnicity remained statistically significant in multivariable models that controlled for insurance type. We conclude that many of the sociodemographic factors associated with the digital divide do not present barriers to accessing context-specific patient education information once in the portal. In fact, this type of highly convenient plain-language patient education may provide particular value to patients in traditionally disadvantaged groups.

BACKGROUND
Providing patients with access to their medical records became a national priority with the federal electronic health record incentive program (the “meaningful use” program) mandating various forms of electronic release of data to patients (1-3). To provide patients with records access, healthcare institutions typically offer patient portals, web-based products that give patients access to data from the electronic health record (EHR) (4-7). However, because medical records are created for medical professionals, they contain technical vocabulary that tends to be unfamiliar to patients, including medication names, diagnoses, and procedures. This vocabulary is particularly challenging for patients with low levels of health literacy (8-10), which has been defined as the ability to obtain, understand, and communicate about health-related information to make informed decisions (10). Low health literacy is associated with increased rates of hospitalization, lower use of preventive care, and among the elderly, worse overall health status and higher mortality (10-12). One-quarter to one-third of Americans are estimated to have limited health literacy, and it is more prevalent among racial and ethnic minorities, the elderly, and those living in poverty (12-14). As a result, many of the patients most in need of information about their health are among those least likely to be able to use it (15).

One potential solution is to provide plain-language explanations of difficult medical terms in context. A precedent is found in the infobutton, a clinician-facing innovation that provides context-specific information in the EHR, which has been shown to help clinicians answer clinical questions quickly at the point of care (16). To provide infobutton-like resources for patients, the National Library of Medicine recently collaborated with EHR vendor Epic Systems Inc. and the Institute for Family Health (IFH), a federally qualified health center. The 3 collaborating institutions developed a service hyperlinking vocabulary terms in the Epic portal with MedlinePlus, the National Library of Medicine’s free online patient education resource (www.nlm.nih.gov/medlineplus). MedlinePlus is an encyclopedia of plain-language explanations of medical diagnoses, medications, and other medical terms, designed to be at the 8th-grade reading level. The service, MedlinePlus Connect, automatically renders medical terms in the electronic patient portal as active hyperlinks. Notably, by hyperlinking unfamiliar terms directly to related information in MedlinePlus, this service helps patients access relevant information without the additional barrier of using a search engine, and without encountering the irrelevant or low quality results that might be produced by a self-directed online search. The service, first implemented at IFH, is now available as a web service and has been implemented in other electronic health record and electronic patient portal systems across the country (https://www.nlm.nih.gov/medlineplus/connect/service.html).

In this study, we sought to evaluate uptake of the innovation by tracking use of the vocabulary hyperlinks by patients over the first 3 years of the service. We hypothesized that use of these vocabulary hyperlinks would be lower among the disadvantaged patient groups affected by the “digital divide” (17).
METHODS

Setting: The Institute for Family Health is a federally qualified health center with 18 sites in Manhattan and the Bronx, as well as clinics in rural areas and smaller towns to the north of the city in the Hudson Valley. Almost all physicians are family practice physicians. An early adopter of EHRs, IFH has offered its patients an English-language electronic patient portal (MyChart) since 2007, with the Spanish-language version added in 2011.

When patients log into their portal account, they can see the conditions listed on their EHR problem list as well as the billing codes from each encounter. Some of these medical conditions are rendered directly as ICD terminology, and a subset are rendered as “patient-friendly” equivalents (e.g., “strep throat” instead of “streptococcal infection of the throat”).

With MedlinePlus Connect (MPC), all of these medical conditions and procedures are automatically rendered as active hyperlinks. Each hyperlink contains the ICD code associated with the medical term, and automatically links to the most relevant MedlinePlus information, which is also indexed by ICD code. Data collection for this study was conducted before IFH transitioned to ICD-10, so all codes were ICD-9.

Study design and data collection: For this retrospective cohort study, all adult patients with a visit to an IFH site between February 2011 and February 2014 were eligible. The primary outcomes were frequency of patient portal access and frequency of use of MPC. For the predictor variables, deidentified reports of patient data were generated from the Epic database and included age, gender, race, ethnicity, insurance status, preferred language, location, number of clinic visits, and encounter diagnoses (during the course of this study, these were ICD-9 codes). We applied the Johns Hopkins Adjusted Clinical Groups algorithm to the ICD-9 codes to classify patients by number of chronic conditions. We attributed each patient to the provider he or she saw the most frequently (his/her “preferred” provider). As an indicator of the provider’s IFH-specific workload, we produced a count of all patients seen by that provider. The project was approved by the IRBs of Weill Cornell and the Institute for Family Health.

Analysis: Patient characteristics were summarized with descriptive statistics, and assessed for association with (A) portal use and (B) MPC use in bivariate analyses using chi-squared tests or independent sample t-tests as appropriate. We constructed multivariable logistic regression models for MPC use among patient portal users only. Model 1 incorporated variables that were significant at .05 in the bivariate analyses (with the exception of number of chronic conditions, which was so strongly correlated with number of encounters that it raised collinearity concerns and was omitted). Region was not included as it was not a significant predictor in bivariate analyses. However, because of the marked differences in the demographic profile of the 3 regions, we also constructed exploratory...
multivariable models stratified by region. These models (not shown) revealed a significant interaction between age and region, with age having different effects on MPC use in different regions. We therefore constructed multivariable Model 2 containing all variables included in Model 1 plus region and a region × age interaction term. After confirming that the Akaike information criterion (AIC) for Model 2 was smaller than the AIC for Model 1 suggesting improved fit, we selected Model 2 as the final model for the current paper.

For exploratory purposes, we also computed a “popularity” index for each vocabulary term that was clicked by the patients. This “popularity” index was the number of patients who clicked the term divided by the number of patients in whose medical record the term appeared. To simplify the graph (figure 1), we reduced the number of vocabulary terms by running the ICDs through the AHRQ/HCUP Clinical Classifications grouper, which groups closely related ICD-9 codes into clinical conditions (e.g., diabetes, endometriosis, etc.).

RESULTS

There were 129,738 adult patients with one or more clinical visits in the study period; 30,692 (24%) had portal accounts. A total of 12,877 (42% of portal users, or 10% of all patients) explored one or more MedlinePlus Connect links. Patients who used the vocabulary hyperlinks clicked a median of 2.0 times (maximum: 97) and a median of 2.0 different terms (maximum, 24).

Table 1: Characteristics of MedlinePlus Connect (MPC) users and nonusers

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total</th>
<th>Portal users</th>
<th>MPC users as % of portal users</th>
<th>MPC non-users as % of portal users</th>
<th>P</th>
<th>MPC users as % of all patients in category</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>129,738 (100%)</td>
<td>30,692 (23.7%)</td>
<td>12,877 (42.0%)</td>
<td>17815 (58.0%)</td>
<td>12877 (9.9%)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>24051 (18.5%)</td>
<td>6195 (25.8%)</td>
<td>2436 (39.3%)</td>
<td>3759 (60.7%)</td>
<td>&lt;.001</td>
<td>10.1%</td>
</tr>
<tr>
<td>25-44</td>
<td>56414 (43.5%)</td>
<td>15702 (27.8%)</td>
<td>6608 (42.1%)</td>
<td>9094 (57.9%)</td>
<td>11.7%</td>
<td></td>
</tr>
<tr>
<td>45-64</td>
<td>38921 (30.0%)</td>
<td>7526 (19.3%)</td>
<td>3338 (44.4%)</td>
<td>4188 (55.7%)</td>
<td>8.6%</td>
<td></td>
</tr>
<tr>
<td>65+</td>
<td>10352 (8.0%)</td>
<td>1269 (12.3%)</td>
<td>495 (39.0%)</td>
<td>774 (60.4%)</td>
<td>4.8%</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>78698 (60.7%)</td>
<td>20708 (26.3%)</td>
<td>9034 (43.6%)</td>
<td>11674 (56.4%)</td>
<td>&lt;.001</td>
<td>11.5%</td>
</tr>
<tr>
<td>Men</td>
<td>51035 (39.3%)</td>
<td>9984 (19.6%)</td>
<td>3843 (38.5%)</td>
<td>6114 (61.5%)</td>
<td>7.5%</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>29774 (23.0%)</td>
<td>6752 (22.7%)</td>
<td>2910 (43.1%)</td>
<td>3842 (56.9%)</td>
<td>&lt;.001</td>
<td>9.8%</td>
</tr>
<tr>
<td>White</td>
<td>46417 (35.8%)</td>
<td>11670 (25.1%)</td>
<td>4701 (40.3%)</td>
<td>6969 (59.7%)</td>
<td>10.1%</td>
<td></td>
</tr>
<tr>
<td>All other</td>
<td>33224 (25.6%)</td>
<td>7926 (23.9%)</td>
<td>3426 (43.2%)</td>
<td>4500 (56.8%)</td>
<td>10.3%</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>20323 (15.7%)</td>
<td>4344 (21.4%)</td>
<td>1840 (42.4%)</td>
<td>2504 (57.6%)</td>
<td>9.1%</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latino, prefers Spanish</td>
<td>10080 (7.8%)</td>
<td>1125 (11.2%)</td>
<td>363 (32.3%)</td>
<td>762 (67.7%)</td>
<td>&lt;.001</td>
<td>3.6%</td>
</tr>
<tr>
<td>Latino, not prefer Spanish</td>
<td>25002 (19.3%)</td>
<td>7116 (28.5%)</td>
<td>3213 (45.2%)</td>
<td>3903 (54.9%)</td>
<td>12.9%</td>
<td></td>
</tr>
<tr>
<td>Not Latino</td>
<td>79141 (61.0%)</td>
<td>19517 (24.7%)</td>
<td>8084 (41.4%)</td>
<td>11433 (58.6%)</td>
<td>10.2%</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>15515 (12.0%)</td>
<td>2934 (18.9%)</td>
<td>1217 (41.5%)</td>
<td>1717 (58.5%)</td>
<td>7.8%</td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>38885 (30.0%)</td>
<td>13035 (33.5%)</td>
<td>5567 (42.7%)</td>
<td>7468 (57.3%)</td>
<td>&lt;.001</td>
<td>14.3%</td>
</tr>
<tr>
<td>Medicaid</td>
<td>39706 (30.6%)</td>
<td>9083 (22.9%)</td>
<td>3959 (43.6%)</td>
<td>5124 (56.4%)</td>
<td>10.0%</td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>28695 (22.1%)</td>
<td>4395 (15.3%)</td>
<td>1572 (35.8%)</td>
<td>2823 (64.2%)</td>
<td>5.5%</td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>15478 (11.9%)</td>
<td>2266 (14.6%)</td>
<td>976 (43.1%)</td>
<td>1290 (56.9%)</td>
<td>6.3%</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>2474 (1.9%)</td>
<td>427 (17.3%)</td>
<td>174 (40.8%)</td>
<td>253 (59.2%)</td>
<td>7.0%</td>
<td></td>
</tr>
<tr>
<td>Other public or dual</td>
<td>4500 (3.5%)</td>
<td>1486 (33.0%)</td>
<td>629 (42.3%)</td>
<td>857 (57.7%)</td>
<td>14.0%</td>
<td></td>
</tr>
<tr>
<td>Encounters &gt;3</td>
<td>68850 (53.1%)</td>
<td>21525 (31.3%)</td>
<td>10182 (47.3%)</td>
<td>11340 (52.7%)</td>
<td>&lt;.001</td>
<td>14.8%</td>
</tr>
</tbody>
</table>

Chronic conditions

| Region         |       |              |                              |                                   |   |                                               |
| Hudson Valley  | 38016 (29.3%) | 5344 (14.1%) | 2272 (42.5%) | 3072 (57.5%) | .60 | 6.0% |
| Bronx          | 24631 (19.0%) | 6486 (26.3%) | 2729 (42.1%) | 3757 (57.9%) | 11.1% |
| Manhattan      | 67090 (51.7%) | 18862 (28.1%) | 7876 (41.8%) | 10986 (58.2%) | 11.7% |
| Workload       |       |              |                              |                                   |   |                                               |
| < 794 patients a year | 32190 (24.8%) | 6613 (20.5%) | 2679 (40.5%) | 3934 (59.5%) | <.001 | 8.3% |
| 794 – 1715     | 32276 (24.9%) | 8419 (26.1%) | 3540 (42.1%) | 4879 (58.0%) | 11.0% |
| "preferred" 1716 – 2714 | 32130 (24.8%) | 7856 (24.5%) | 3472 (44.2%) | 4384 (55.8%) | 10.8% |
| 2715 or more   | 33142 (25.6%) | 7804 (23.6%) | 3186 (40.8%) | 4618 (59.2%) | 9.6% |
Sociodemographic predictors of use of MedlinePlus Connect (bivariate)

As demonstrated in Table 1, use of the informational hyperlinks was associated with socioeconomic characteristics but not in the hypothesized direction. (Throughout Table 1, it is important to note that because of the large sample size, very small differences were sometimes statistically significant even if they are unlikely to be clinically significant.) Black patients were more likely to use MPC than white ones (43% compared to 40%), and Latino patients more likely to use the resource than non-Latinos (43% compared to 41%). When Latinos were further subdivided by their preferred language, it became clear that it was the English-preferring Latinos who were driving the use of this resource. Finally, patients covered by Medicaid were more likely to use MPC than privately insured ones (44% versus 43%).

Clinical predictors of use of MedlinePlus Connect (bivariate)

Use of MPC was associated with clinical characteristics in the expected direction. Women, patients with more encounters, and patients with more chronic conditions were more likely to use the resource. MedlinePlus Connect users were very slightly older than nonusers, due primarily to greater representation in the middle-aged (45-64 years) age bracket.

Health system predictors of use of MedlinePlus Connect (bivariate)

Patients of the least busy and most busy providers were least likely to use MPC. There were no differences by region.

Table 2: Multivariable model of MPC use by sociodemographic characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Level</th>
<th>Adjusted Odds Ratio</th>
<th>CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>(per 1-year increase)</td>
<td>1.004</td>
<td>1.002</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sex</td>
<td>Women</td>
<td>1.167</td>
<td>1.109</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>Black</td>
<td>1.100</td>
<td>1.027</td>
<td>.007</td>
</tr>
<tr>
<td></td>
<td>All other</td>
<td>1.143</td>
<td>1.056</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>1.106</td>
<td>1.010</td>
<td>.02</td>
</tr>
<tr>
<td>Ethnicty with preferred language</td>
<td>Latino, does not prefer Spanish</td>
<td>1.077</td>
<td>1.001</td>
<td>.045</td>
</tr>
<tr>
<td></td>
<td>Latino, prefers Spanish</td>
<td>0.607</td>
<td>0.525</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Unknown ethnicity</td>
<td>0.993</td>
<td>0.900</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>Not Latino</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance</td>
<td>Medicaid</td>
<td>0.897</td>
<td>0.847</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Medicare</td>
<td>0.865</td>
<td>0.782</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>Other Public or dual</td>
<td>0.960</td>
<td>0.859</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>Uninsured</td>
<td>0.766</td>
<td>0.711</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>0.888</td>
<td>0.726</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;3</td>
<td>2.164</td>
<td>2.049</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>&lt; 794 patients a year</td>
<td>1.093</td>
<td>1.022</td>
<td>.01</td>
</tr>
<tr>
<td>Provider workload</td>
<td>794 – 1715</td>
<td>1.173</td>
<td>1.094</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>1716 – 2714</td>
<td>1.034</td>
<td>0.965</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>&gt;2715 or more</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Hudson Valley</td>
<td>1.019</td>
<td>0.841</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td>Bronx</td>
<td>1.379</td>
<td>1.161</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Manhattan</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hudson Valley*Age</td>
<td>0.997</td>
<td>0.993</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>Bronx*Age</td>
<td>0.988</td>
<td>0.983</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Manhattan*Age</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multivariable model of MedlinePlus Connect use

The final multivariable model included: age, gender, race, ethnicity with preferred language, insurance type, number of encounters, workload of preferred provider, region, and the region x age interaction term (Table 2). In this model, use of the informational hyperlinks remained positively associated associated with older age, female sex, black race,
English-preference Latinos, having private insurance, more clinical encounters, and provider workload (with MPC use being more common forms patients of less busy providers).

Interestingly, when we controlled for race and ethnicity in this multivariable model, patients with Medicaid were less likely to use the MPC resource. The positive relationship in the bivariate analysis was therefore due to the black and Latino patients who were using the resource. The multivariable results suggest that within each racial and ethnic category, patients with Medicaid (the lower income patients) were somewhat less likely to use the MPC resource.

As noted in the methods, we included the interaction term between region and age because we saw different age-related patterns in preliminary bivariate models that were stratified by region (not shown). These models showed that older patients were more likely than younger patients to use MPC in Manhattan but not in other regions. To confirm this finding in the multivariate models, we included an interaction term between region and age, and confirmed that it was statistically significant in the expected direction. In other words, as Manhattanites became older, they were more likely to use MPC, but this pattern was much weaker in the Bronx and not found at all in the Hudson Valley. We attribute this finding to the different demographic patterns in these regions. In Manhattan, the older population contains quite a number of people who are well-educated and have relatively high socioeconomic status, so they might be more likely to do research on their health conditions. By contrast, the Bronx and the small towns where IFH has a presence in the Hudson Valley are much less affluent, and elderly patients in those areas are likely to be less well-educated and less affluent members of minority groups.

**Frequently explored terms**

Diagnosis “popularity” (or proportion of times a diagnosis term was clicked) was fairly constant at an average of 14.5% regardless of how common the diagnosis was (Figure 2). “Popularity” appeared high for some of the rare conditions, but this was in fact an artifact of the rarity of the condition. For example, *multiple myeloma* had a very high popularity rating of 56%, but this was because only 9 patients had the condition and 5 of them clicked upon it.

![Figure 2: Each dot represents a diagnosis. Data labels are placed on a small number of “popular” terms for illustrative purposes.](image-url)
As Figure 2 shows, there was no obvious explanation for term “popularity.” To further explore “popularity,” we retrieved the 10 terms that were clicked by the largest numbers of patients (Table 3). No single unifying principle is immediately apparent to explain why these terms were frequently clicked. It is possible that in some cases, it is because the terms were unfamiliar and technical (“hyperlipidemia,” “unspecified essential hypertension”), prompting patients to seek a definition. It is also possible that people were more likely to click on stigmatized or contested diagnoses (e.g., “obesity,” “depression,” “anxiety”). In other cases, perhaps patients clicked when they were motivated to take action and do more research (“nicotine dependence,” which is likely to appear in the record of the patient already being treated for smoking cessation). A better understanding of why patients seek more information about certain conditions and not others would be a fruitful avenue for future quantitative and qualitative work.

Table 3: Top 10 most explored terms

<table>
<thead>
<tr>
<th>ICD9 code</th>
<th>Diagnosis term</th>
<th>Number of patients who clicked this term</th>
<th>Number of portal users with this diagnosis</th>
<th>Clickers as percent of portal users with diagnosis (&quot;popularity&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V70.0</td>
<td>General adult medical examination</td>
<td>4638</td>
<td>21,940</td>
<td>21.1</td>
</tr>
<tr>
<td>278.00</td>
<td>Obesity</td>
<td>929</td>
<td>3896</td>
<td>23.8</td>
</tr>
<tr>
<td>V25.9</td>
<td>Contraceptive management</td>
<td>895</td>
<td>5471</td>
<td>16.4</td>
</tr>
<tr>
<td>272.4</td>
<td>Hyperlipidemia</td>
<td>619</td>
<td>3560</td>
<td>17.4</td>
</tr>
<tr>
<td>401.9</td>
<td>Unspecified essential hypertension</td>
<td>407</td>
<td>3027</td>
<td>13.5</td>
</tr>
<tr>
<td>311</td>
<td>Depression</td>
<td>358</td>
<td>2216</td>
<td>16.2</td>
</tr>
<tr>
<td>V62.9</td>
<td>Psychosocial circumstances</td>
<td>362</td>
<td>1560</td>
<td>23.2</td>
</tr>
<tr>
<td>300.00</td>
<td>Anxiety</td>
<td>317</td>
<td>1962</td>
<td>16.2</td>
</tr>
<tr>
<td>305.1</td>
<td>Nicotine dependence</td>
<td>306</td>
<td>2915</td>
<td>10.5</td>
</tr>
<tr>
<td>401.1</td>
<td>Benign essential hypertension</td>
<td>242</td>
<td>1749</td>
<td>13.8</td>
</tr>
</tbody>
</table>

DISCUSSION

Although providing patients with access to their medical records could help them better understand their health and healthcare, patients who do not understand the medical language found in these medical records are unlikely to benefit. Our study shows that a plain-language encyclopedia of medical terms, hyperlinked directly to the unfamiliar term, is frequently explored by patients with access to their medical records via an electronic patient portal. About 10% of all patients in this safety net population used the informational hyperlinks. The hyperlinks were appropriately used most by those with the greatest needs for medical information, i.e., those with more medical conditions and visits.

The sociodemographic analysis held some surprises. We and others have expressed concern that because of disparities in computer access as well as disparities in health literacy, any beneficial effects of patient portals are likely to be unequally distributed (19). However, counter to the main hypothesis of the current study, we found positive associations between MPC use, black race, being an English-speaking Latino patient, and being covered by Medicaid. The positive associations with black race and Latino ethnicity remained statistically significant in multivariable models that controlled for insurance type. However, in the multivariable model, Medicaid coverage developed a negative relationship with MedlinePlus Connect use. This can be interpreted to mean that Blacks and English-speaking Latinos within each insurance category are more likely to use the informational hyperlinks, although Medicaid (low income) patients are in general less likely to do so.

Latino patients who indicated that they prefer using Spanish in their clinical encounters were markedly less likely to use the resource. This was despite the availability of a Spanish-language portal and Spanish-language informational resources.

Overall, this study suggests that some of the factors associated with the digital divide which created barriers to accessing the electronic portal were not barriers to using the MedlinePlus Connect informational hyperlinks. However, the language barrier remains indicating the need for additional outreach and services to ensure Spanish-speaking patients have high-quality information about their health.

There are a number of potential explanations for why patients might click terms. One explanation is that patients with low health literacy (9, 11) would be less likely to understand a particular medical term and might be
more likely to click to find out more. Unfortunately, because we were using EHR data, we had no assessments of patient health literacy, so we cannot conclude definitively that patients with low health literacy were the ones who were clicking on the hyperlinks. However, the prevalence of low health literacy is known to be higher among minority and low income patients (9, 11), and our findings demonstrate that these groups were particularly likely to use the MPC links. If low health literacy is the explanation, this would suggest that the MedlinePlus Connect links are providing particular value to patients who have less familiarity with medical vocabulary. This could provide an explanation for certain of the most “popular” terms in Figure 2, such as aneurysm and endometriosis, multisyllabic words likely to be unfamiliar to lay audiences.

Other reasons why a patient might click a term could include being highly concerned about a diagnosis, needing more explanation than what the provider discussed, or feeling uncomfortable having an extended conversation with the provider. For example, it is possible that patients with sensitive, stigmatized, or contested diagnoses (20-22) might be interested in finding out more in private after the encounter, when visiting the patient portal. These could explain the high click rates for potentially sensitive terms such as obesity and psychosocial circumstances in Table 3, and for highly concerning diagnoses such as sickle cell in Figure 2.

**Limitations**

This was a study at a single center providing safety net care, and all patients used a single commercially available patient portal product, so generalizability to other products and populations is unknown. Analyses were limited to data available in log files or the EHR, and therefore no direct measurements of health literacy, access to computers, or patient perceptions were available. This study did not include any measurement of healthcare outcomes, so it is not known whether access to this information actually helped patients manage their own health or their healthcare.

**Conclusions**

Context-specific hyperlinks that provided plain-language explanations of medical vocabulary were heavily used by electronic patient portal account holders at a federally qualified health center, especially those with multiple office visits and clinical conditions. Socioeconomic factors that presented barriers to access to the portal (such as race and poverty) did not pose barriers to use of the vocabulary hyperlinks within the portal. In fact, hyperlink users were more likely to be black or English-speaking Latino than were non-hyperlink users. We conclude that context-specific medical vocabulary hyperlinks are valuable to patients with the greatest information needs because of greater use of healthcare services, and that they are of particular use to patients in disadvantaged populations.

**ACKNOWLEDGMENTS**

Dr. Ancker is funded by AHRQ K01 HS021531. This evaluation also received funding from the National Library Medicine. However, neither funder played any role in study design, data analysis or interpretation, or review or approval of the manuscript before publication.

The MedlinePlus Connect service was developed by the National Library of Medicine, the Institute for Family Health, and Epic Systems Inc. The authors thank Rob Logan of the NLM for significant contributions to the MPC implementation.
REFERENCES
Studying Readiness for Clinical Decision Support for Worker Health Using the Rapid Assessment Process and Mixed Methods Interviews

Joan S. Ash, Ph.D., M.L.S., M.B.A.¹, Dian Chase, Ph.D., R.N., F.N.P.¹, Jane F. Wiesen, Ph.D.¹, Elizabeth V. Murphy, M.D., M.P.H.¹, Stacey Marovich, M.S., M.H.I.²

¹ Oregon Health & Science University, Portland, OR, USA
² CACI, Inc.

Abstract

To determine how the Rapid Assessment Process (RAP) can be adapted to evaluate the readiness of primary care clinics for acceptance and use of computerized clinical decision support (CDS) related to clinical management of working patients, we used a unique blend of ethnographic methods for gathering data. First, knowledge resources, which were prototypes of CDS content areas (diabetes, lower back pain, and asthma) containing evidence-based information, decision logic, scenarios and examples of use, were developed by subject matter experts. A team of RAP researchers then visited five clinic settings to identify barriers and facilitators to implementing CDS about the health of workers in general and the knowledge resources specifically. Methods included observations, semi-structured qualitative interviews and graphic elicitation interviews about the knowledge resources. We used both template and grounded hermeneutic approaches to data analysis. Preliminary results indicate that the methods succeeded in generating specific actionable recommendations for CDS design.

Introduction

The majority of adults in the U.S. work, and on average they spend more than half their waking hours at work.¹ Therefore, primary care providers treating adults encounter many situations where the patient’s work environment affects their health and/or the management of their health conditions. And although primary care providers most often are the first to see patients with medical issues such as asthma symptoms that may be caused by workplace exposures, they do not routinely ask patients about their work.²⁻⁶ Occupational health physicians have developed evidence-based guidelines for helping to manage many such patients, but primary care providers are rarely aware of their existence.⁷ Though computerized clinical decision support (CDS) holds potential for increasing awareness and providing guidance in the care of working patients, it must be developed and implemented to fit the context and workflow of those who would benefit from having the information.⁸

This qualitative evaluation project was part of a larger project funded by the National Institute for Occupational Safety and Health (NIOSH) of the Centers for Disease Control and Prevention. The NIOSH project is designing, developing, and pilot testing clinical decision support (CDS) for the health of working patients in primary care outpatient settings. The first step in this project was the development of three knowledge resources (KRs) containing evidence-based information, decision logic, scenarios and examples of use. The KRs were prepared by three subject matter expert (SME) groups for three topics that are related to the health of patients who work and that were considered especially pertinent to a primary care practice. The SME groups were guided through the guideline and KR development process by an informatician with expertise in these procedures (Dr. Richard N. Shiffman). The three KRs focused on dealing with work environment factors that impact the management of a chronic disease (diabetes), guidance for return-to-work (RTW) after lower back pain diagnosis not related to work, and diagnosis and management of work-related/work-exacerbated asthma. The goal of the qualitative study was to identify the barriers and facilitators related to CDS for the clinical management of working patients in a variety of primary care settings, including assessment of the technical and organizational feasibility of implementing the CDS represented by each KR. Having used the Rapid Assessment Process (RAP) in the past, the research team members considered it most appropriate for this study.
RAP is a way of gathering, analyzing, and interpreting ethnographic data that is both effective and efficient. The process is expedited through the consistent use of structured tools, which are developed for each individual study and consolidated into a field manual. RAP can only be conducted by teams that include those inside the organization as well as outside researchers. RAP also provides feedback to internal stakeholders. In summary, RAP “depends heavily on triangulation of data from different sources. A field manual is developed prior to the study and generally includes 1) site inventory profiles, 2) observation guides, 3) interview question guides, and 4) rapid survey instruments.” As with traditional ethnography, methodological approaches must be tailored specifically to the needs of the project.

The RAP methodology has mainly been applied to assessing clinical systems that have already been implemented and within which CDS is embedded, but for this project we aimed to adapt it for assessing both the clinical context and the KR prototypes prior to full development and implementation. Because we needed direct comments about each of the three KRs, we explored using graphic elicitation interviewing techniques in addition to our classical semi-structured interview methods. Graphic elicitation interviewing involves asking the subject to look at a graphical artifact to stimulate discussion, providing “contemplative verbal responses.” Umoquit et al. have compared the use of graphic elicitation, which involves use of a visual aid available to both the interviewer and interviewee, with participatory diagramming, in which the interviewee is tasked with developing a graphic. They found that graphic elicitation was best for focusing on exactly what the researchers were interested in and for producing thoughtful verbal comments on researcher-identified issues. Participatory diagramming produced fewer detailed comments about those issues but it encouraged creativity and new ideas. For our purposes, to gather verbal commentaries about specific aspects of the KRs, we selected graphic elicitation as our method of choice along with semi-structured interviewing.

For the study described here, our research question was: How can RAP be further adapted for assessment of the readiness of the organizational and technological context for CDS and, in addition, for evaluation of the CDS content?

Methods

Selection of methodological approaches

Because of the timeline for the larger project and the need to request human subjects approval from NIOSH, OHSU and five different organizations serving as sites, we only had five months within which to conduct site visits, analyze data, and report results. We fortunately had a large enough team of trained researchers so that some visits by sub-teams could take place at the same time. Our most difficult decisions concerned maximizing the use of our interviews so that we optimized the use of interview time, learned about context, and also received detailed input about the KRs. We opted to conduct two-part interviews using two different interviewing strategies.

For gathering input on the KRs, we wanted interviewees to evaluate a CDS concept with which they may not have been familiar, especially if they were non-clinicians. Further, we needed them to react in an abstract way because the CDS was not yet built. We decided we needed a variety of tools to help us describe the CDS ideas to our subjects. We designed two artifacts in addition to the extremely detailed documents developed by each of the three SME groups: a brief easily understood one-paragraph description of the CDS and a flowchart showing what the CDS would do. We planned to get general assessments of the usefulness of the CDS from non-clinical and non-technical interviewees using these less detailed descriptions. For others, we used different artifacts depending on the expertise of the interviewee.

We also decided that interviews alone would not be sufficient. To learn about the workflow within which the proposed CDS might best fit, we opted to conduct observations. This form of methodological triangulation serves to verify whether or not workflow described during interviews is accurate. Researchers shadowed clinic staff and clinicians and observed work activities throughout each clinic while writing detailed field notes. These observations also provided an opportunity to see how patient work information was recorded in the EHR.
Development of a Field Manual

Our multidisciplinary team of nine occupational health experts and informaticians from four different U.S. locations convened in Portland, OR for three days of training and planning. The training had RAP as its focus since some team members were familiar with ethnographic techniques but not with RAP in particular. Decisions were made together as we developed the field manual for this study. For our field manual, we put together a “site inventory profile,” which was a checklist of kinds of CDS and other related factors. We gathered some information prior to our visit to help us develop our interview questions and lists of foci for observations and we also gathered some site inventory profile information on site. We conducted phone discussions with a contact person at each site to learn more about the site and to mutually identify individuals to interview and observe. As is our practice, our field manual included a master interview guide with all of the questions we wanted to have answered. For each individual, we selected the most appropriate questions depending on his or her role. As with other field manuals we have prepared, the field manual also included a fieldwork observation guide for use in writing field notes. It briefly noted foci for fieldwork at that particular site. Finally, we included the three different artifacts for each of the three KRs, which would help us with the graphic elicitation interviews.

Human Subjects Protection

This study was approved by the Institutional Review Boards (IRBs) at NIOSH, OHSU, and two of our study sites. IRB review and/or official determinations (if a site did not have an IRB, it had other mechanisms for reviewing our protocol) were also obtained from the other three sites.

Site selection

We deliberately selected geographically-diverse sites and both large and small organizations so that we could observe whether these differences might influence the ability to customize, implement, and gain user satisfaction with the proposed CDS; and to get a sense of capability, and applicability of our findings across a variety of clinical settings. We also purposely selected sites that varied in organizational structure and had different electronic health record systems (EHRs).

We conducted this qualitative assessment across several clinics within two Federally Qualified Health Centers (FQHC) using the NextGen EHR, two organizations of very different sizes that use Epic, and one very large, complex organization that uses AllScripts. Sites were located in California, Mississippi, New York, Massachusetts, and Ohio.

Subject selection

With the assistance of an inside contact person at each site, we selected individuals representing different roles on health care teams, including physicians, nurses, pharmacists, medical assistants, social workers, care coordinators, educators, clerical and administrative staff, and information technology and CDS specialists.

Data collection

The interviews explored present CDS usage in general, work information in commercial EHRs, capability for implementing new CDS, information needs related to managing care of working patients, workflow, and assessment of the three KRs.

We developed tailored interview guides for individuals depending on their roles in the organization. During the first half of the interview, we asked general questions about CDS, information to help manage care of working patients (such as availability of the patient’s work information in the EHR), technical issues, and workflow (Table 1). A sample of some questions asked of information technology-related staff is given in Table 2. Figure 1 shows an example of one of the three flowcharts used in the second part of the interview as a graphic elicitation tool. For the second part of each interview, we asked specific questions about the KRs (see Table 3 for an example), assisted by the artifacts (text description, flowchart, or full KR) selected as most appropriate. Subjects representing all roles were first shown the paragraph summarizing each of the proposed CDS modules. Subsequently, depending on the role of the interviewee and his or her level of understanding and use of CDS, we used the graphic elicitation technique. We often showed the flowchart in addition to the paragraph and on rare occasions we showed the full KRs.
We also adjusted our questioning during the interview so that we could probe intriguing topics. Two interviewers attended each interview, with a few exceptions due to scheduling. We tried to include one occupational health specialist and one informatics specialist on each team so that each team had the expertise to probe either clinical or technical answers given by interviewees. One interviewer was the official interviewer and the other remained silent until appropriate times when he or she was allowed to ask questions during the first part of the interview. In this way, the process remains true to recommended semi-structured interview techniques (e.g. it does not become a conversation), but the assistant interviewers have an opportunity to ask follow-up questions in their particular areas of expertise. Another important task for the assistant interviewer was to write field notes during the interview. They noted nonverbal interactions and areas to further explore as well as what the person was saying. For the second part of the interview, when we used the graphic elicitation technique to discuss the flowcharts and/or KRs, both of the interviewers played a more active role since more of a dialogue and question and answer interchange was necessary. All interviews were recorded and transcribed.

Table 1. Topics in interview guides

<table>
<thead>
<tr>
<th>1. Clinical interviewee question areas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part 1</strong></td>
</tr>
<tr>
<td>Background and role of interviewee</td>
</tr>
<tr>
<td>Your work patterns</td>
</tr>
<tr>
<td>About CDS</td>
</tr>
<tr>
<td>About CDS for clinical management of patients who work</td>
</tr>
<tr>
<td><strong>Part 2</strong></td>
</tr>
<tr>
<td>The three knowledge resources: how useful, who should be involved in asking questions and educating, and where each fits into the workflow</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Informatics and IT question areas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part 1</strong></td>
</tr>
<tr>
<td>Background and role of interviewee</td>
</tr>
<tr>
<td>Clinician work patterns and clinic workflows</td>
</tr>
<tr>
<td>About your EHR</td>
</tr>
<tr>
<td>About training, support, and customization</td>
</tr>
<tr>
<td>About CDS</td>
</tr>
<tr>
<td>About CDS for clinical management of patients who work</td>
</tr>
<tr>
<td><strong>Part 2</strong></td>
</tr>
<tr>
<td>The three knowledge resources: how useful, who should be involved in asking questions and educating, and where each fits into the workflow</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Management and staff question areas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part 1</strong></td>
</tr>
<tr>
<td>Background and role of interviewee</td>
</tr>
<tr>
<td>Your work patterns</td>
</tr>
<tr>
<td>About CDS</td>
</tr>
<tr>
<td>About CDS for clinical management of patients who work</td>
</tr>
<tr>
<td><strong>Part 2</strong></td>
</tr>
<tr>
<td>The three knowledge resources: how useful, who should be involved in asking questions and educating, and where each fits into the workflow</td>
</tr>
</tbody>
</table>

During our observations, we observed all activities throughout the clinic, from patient registration through patients leaving the clinic, so that we could trace workflow in a general sense. We shadowed individual
providers and staff members while they were interacting with patients and also when they were performing other duties. We also conducted informal interviews with those we were observing when there were opportunities.

Table 2. Sample of questions asked of technical experts

<table>
<thead>
<tr>
<th>4. About work patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Could you briefly describe the work pattern or different work patterns of clinicians in your clinic—what are their days like?</td>
</tr>
<tr>
<td>What do they do before, during, and after patient visits?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5. About your EHR system</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see here that your organization uses X EHR system.</td>
</tr>
<tr>
<td>Can you give an overview of the configuration of your EHR system? Are there separate interfaces/modules for the registration vs. billing vs. patient chart?</td>
</tr>
<tr>
<td>(if there are separate modules) Do you know how/if data are shared among the different modules?</td>
</tr>
<tr>
<td>How are data entered into your EHR system? (probe -- are any data entered directly by patients via a tablet, kiosk, or patient portal/PHR system?)</td>
</tr>
<tr>
<td>Could you describe how the process works for requesting, testing, and deploying modifications to your EHR system?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. About clinical decision support</th>
</tr>
</thead>
<tbody>
<tr>
<td>We define decision support to include alerts, reminders, order sets, reference sources, data displays like flow sheets, and documentation templates.</td>
</tr>
<tr>
<td>What kinds of CDS do you have here and what seems to have been most helpful?</td>
</tr>
<tr>
<td>What employment or occupational information is currently captured in your EHR system (such as employer name, occupation/job title, etc.)? Is it in structured form? Where is it in the system?</td>
</tr>
</tbody>
</table>

At the end of each day, the team met to debrief about findings for that day, to share insights from each researcher’s unique perspective, and to plan for the next day. When we had two teams in the field in different geographic locations, we met via teleconference.

Data analysis

Preliminary data analyses took place between interviews. Using notes taken by the assistant interviewer, the researchers briefly reviewed what had been learned so that new questions for upcoming interviews could be developed as needed based on prior interviews. After completion of the site visits, interview notes and transcripts were entered into NVivo, (QSR International, Doncaster, Victoria Australia), a program that facilitates the organization and retrieval of qualitative data for analysis.

We used both grounded hermeneutic and template approaches to data analysis. For in-depth analysis of all data, we used the former, which begins with the words of our interviewees. Our core research team broke into dyads and each dyad was assigned a set number of transcripts. Individuals read the assigned transcripts, noting all recurring or potentially important expressions and key phrases, and the dyads then met to compare and agree on their findings. The researchers next met so each dyad could discuss results with the larger team, which reached consensus on meaning and terminology for themes. One dyad also analyzed and categorized all comments related to the KRs using a template of predetermined terms related to the questions we asked.

The interpretive process was both iterative and flexible. Our team met repeatedly to interpret the results. We wrote a short report of the findings for each site we studied for two reasons: first, we thought the organizations would find them useful, and second, the report was a form of “member checking,” a qualitative technique to further establish trustworthiness of results by asking insiders for feedback.
Results

We conducted five site visits over an 8-week period between July 21st and September 11th, 2015. Team members included two occupational health physicians, three NIOSH staff members (a scientist, a nurse, and an informatician), four informatics faculty members (a social scientist, a nurse practitioner, a physician, and a laboratory specialist), and a project manager. We interviewed 41 clinicians not deeply involved in

Figure 1. Flowchart for use during graphic elicitation interview: diabetes example

---

290
informatics, 23 individuals who were informaticians or information technology specialists, 15 managers or staff members, and four quality improvement specialists, for a total of 83 interviews. We spent a total of 30 hours observing in ten clinics within these five organizations. We believe we reached data saturation in that we reached a point where we were hearing the same answers repeatedly and not learning anything new.

Table 3. Questions for graphic elicitation interviews: diabetes example

<table>
<thead>
<tr>
<th>Part 2 of interviews: CDS Knowledge Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation: An outside panel has come up with three ideas for CDS which might help primary care providers meet patient’s needs more effectively. The three are: using CDS to help identify and manage work related asthma, analyzing the impact of working conditions on refractory diabetes, and providing structured return to work letters for patients with low back pain. We would like to discuss each of them with you.</td>
</tr>
<tr>
<td><strong>[Hand copy of most appropriate artifact to person]</strong></td>
</tr>
<tr>
<td>Interviewee reads:</td>
</tr>
<tr>
<td>Refractory diabetes: A diabetic patient’s working conditions, such working more or different hours or working in hot environments can contribute to hyper or hypoglycemic episodes. Also, for some “safety sensitive” jobs, a worker with impaired cognition due to low blood sugar could be at risk for injury to himself or to others. The CDS would prompt providers to ask specific work-related questions and would generate educational information for the provider and patient based on the responses.</td>
</tr>
<tr>
<td>Ask:</td>
</tr>
<tr>
<td>How useful would this information be to those in your clinic?</td>
</tr>
<tr>
<td>Which clinic personnel are likely to be involved in either gathering patient data or acting on decisions/information like this?</td>
</tr>
<tr>
<td>Who would be asking the required questions?</td>
</tr>
<tr>
<td>Who would/should be educating the patient?</td>
</tr>
<tr>
<td>At what point in the clinic workflow would/should this happen?</td>
</tr>
</tbody>
</table>

Lessons Learned About Methods

Timeline: This was planned as a one-year project, but half of that time was needed to secure the necessary IRB approvals. Although we were able to gather a great deal of information in eight weeks, and to analyze it to produce the needed reports within a three month period after that, this time frame was too short to allow for truly in depth analysis within the contract period. Analysis was still ongoing three months later. Another problem with the short timeline was that, even with nine researchers available to conduct site visits, some had to be done simultaneously and therefore the lead investigator (JA) was not able to attend all visits. In the past, we have found that having the lead investigator attend all visits provided consistency in data gathering (J. Ash, personal observation).

Two-part interview format: The two-part interview strategy worked well most of the time. By progressing from questions about CDS in general to specifics about the KRs, we were able to set the stage so that interviewees knew the context of our questions and were more likely to be comfortable talking with us. We ran out of time during a few of the interviews, which were normally scheduled for a half hour. These occurrences could have indicated that the interviewees were comfortable with talking, but was likely due primarily to trying to cover so much during each interview.

Graphic elicitation interviews: These yielded immensely useful information about the KRs. We always started the second part of the interview by handing the interviewee one paragraph about one KR at a time (or paragraphs about all three KRs if the interviewee preferred seeing them all at once). For the interviewees with more expertise in clinical and/or technical areas, the paragraph started a dialogue, the interviewee asked questions, and the flowchart would then be discussed. This situation made the second
part of the interviews, collecting information about the KRs, especially challenging because the
interviewers not only had to adapt each question to the skill level of the interviewee, but also bring out the
graphical representation of the CDS at the right time. Usually semi-structured ethnographic interviews
using RAP are not dialogues: the interviewee answers questions posed by the interviewer, who tries to do
more listening than talking. However, we found that our subjects could not answer our questions without
the ability to ask interviewers about details of the CDS flow. The graphic elicitation interviews were used
to provide more details simply and quickly, in order to help the interviewees answer our questions and
move back to more of an interview (vs. a dialogue). In this manner, we saved considerable time by only
needing to bring out the full KR documentation a few times when interviewees had very specific questions.

Assistant interviewer: The role of the assistant interviewer was more important in this project than others
we have conducted because the site visits followed one another in such rapid succession that transcripts
from prior visits were not available to help us prepare for the next visit. In fact, it was very important that
the interview notes were detailed enough so that, because of the tight timeline, we were able to write
preliminary reports for each SME group. The reports made recommendations about each KR based on
these field notes without having all of the transcripts completed. Subsequent examination of the transcripts
provided more nuanced information, but did not change the big picture of the information collected.

Subject selection: By observing and interviewing staff in all roles, including clerical staff, we were able to
assess all activities in the clinics and gain a broad picture of clinic workflows and the way patient work
information data are currently collected in the EHRs. Some of these interviewees were unable to provide
much feedback about the KRs, but they all were able to describe their daily work and the way they used the
EHR in detail.

Observations: The notes from the observations provided critical insights into the use of the EHR and CDS.
They allowed us to obtain more detailed information about the workflow, availability of multiple roles
within the clinic, and differences in how these roles are used across the clinics. They also allowed us to
modify or verify our interpretations of what was said in the interviews about workflow and roles. Staff
members often fail to describe their work activities accurately and completely, most likely because they
take many of them for granted and are trying to be succinct during interviews. Observations serve to fill in
the gaps and proved to be another valuable technique for triangulating data.

Lessons Learned About CDS for Clinical Management of Working Patients
Preliminary analysis of the interview and observation data resulted in themes that reflect facilitators or
barriers to development and implementation of the proposed CDS and are indicative of the depth of
knowledge we were able to obtain using RAP. Full results will be reported when analysis is complete, but
analysis of the transcripts and field notes up to this point indicates that there are many facilitators related to
development and acceptance of decision support tools as outlined in the KRs developed by the SME
groups. The most important and somewhat unexpected finding was that it is technically feasible to develop
any of the three proposed CDS tools. There are also a number of barriers to implementation and use. Not
surprisingly, interviewees most consistently noted that the proposed CDS might add to their time burdens.

Responses about the Three Knowledge Resources
From the graphic elicitation interviews we learned that in general the three proposed CDS modules
included some terminology, especially when offering recommendations, that needed explanation. For
example, the term “safety sensitive activity” was not immediately understood and could be changed to
“activity with potential safety risks.” We learned that informatics innovations that facilitate patients
entering data about their work may have potential to relieve some clinic time burdens, that culturally
sensitive patient education materials are needed, and that providers do not necessarily need to be the targets
of the CDS. We were able to gain feedback on each of the KRs that was very detailed, especially when the
interviewee had both clinical and informatics knowledge. As an example, one interviewee with an
informatics background suggested ways to use a service-oriented architecture approach for one of the KRs.
Although we gained useful feedback from medical assistants and other non-provider staff about workflow,
we found that they were less interested in either the flowcharts or KRs.
Discussion

We believe that use of the Rapid Assessment Process is effective for gaining a good deal of information in a short amount of time. For this project, because of circumstances beyond our control, we were forced to work even more rapidly than usual. As others have found when conducting RAP evaluations, flexibility and creativity are needed so that the mix of methods can best fit the situation. Workarounds such as the use of detailed assistant interviewer notes and splitting up the team to conduct simultaneous site visits helped us to stay on schedule, but they are not optimal. On the other hand, we succeeded in gathering a great deal of information about the facilitators and barriers to developing and implementing CDS for the health of working patients in primary care settings based on a combination of interviews and observations.

Enhancing RAP by using graphic elicitation interviews such as those described by Umoquit to assess CDS ideas allowed us to gather rich and specific feedback about the KRs. Like Crilly et al., we found that the artifacts we presented to interviewees served as a reference point and were more effective with many interviewees than simply asking a series of questions. Additionally, like Larkin et al., we found the artifacts most useful for interviewees with the interest and ability to interpret them quickly. The modifications to RAP used for this study might be adopted for other studies of CDS with different content and contexts, though we would recommend a longer timeline.

Recommendations and Conclusion

The use of RAP to evaluate the content and concept of CDS after it has been outlined in a form that is understandable to users but before it has been built can offer a useful description of the perceived value and context within which the CDS can work best. By adding graphic elicitation interviews to the RAP process, researchers can gather detailed feedback for CDS developers and implementers.

Acknowledgements

This project was supported by CDC/NIOSH Contract 200-2015-61837 as part of NORA project #927ZLDN. We would like to thank the following for service on the research team: Sherry Baron, MD, MPH, Genevieve Barkocy Luensman, PhD, Margaret Filios, RN, MSc, Nedra Garrett, MS, Richard N. Shiffman, MD, MCIS, and James McCormack, PhD. We appreciate the help of the following at our sites: Rose H. Goldman, MD, MPH, Laura Brightman, MD, Stacey Curry, MPH, Larry J. Knight, MSA, MSHRM, Deborah Lerner, MD, Herb K. Schultz, Michael Rabovsky, MD, Joseph Conigliaro, MD, MPH, and Nicole Moodie Donoghue. Finally, we would like to acknowledge and thank the clinics and the clinicians and staff at each site for their participation and support of this work.

References

Automatic Generation of Conditional Diagnostic Guidelines

Tyler Baldwin, Ph. D\textsuperscript{1}, Yufan Guo, Ph. D\textsuperscript{1}, Tanveer Syeda-Mahmood, Ph. D\textsuperscript{1}
\textsuperscript{1}IBM Almaden Research, San Jose, CA, USA

Abstract

The diagnostic workup for many diseases can be extraordinarily nuanced, and as such reference material text often contains extensive information regarding when it is appropriate to have a patient undergo a given procedure. In this work we employ a three task pipeline for the extraction of statements indicating the conditions under which a procedure should be performed, given a suspected diagnosis. First, we identify each instance in the text where a procedure is being recommended. Next we examine the context around these recommendations to extract conditional statements that dictate the conditions under which the recommendation holds. Finally, corefering recommendations across the document are linked to produce a full recommendation summary. Results indicate that each underlying task can be performed with above baseline performance, and the output can be used to produce concise recommendation summaries.

Introduction

Understanding which diagnostic procedure should be performed on a patient when presented with a set of symptoms or a suspected diagnosis is a difficult task that is at the heart of the diagnostic process. Often, the details surrounding when a test is appropriate are extraordinarily nuanced, making this a potential area for clinician error. As errors in this regard can put unnecessary burden on the patient and lead to higher costs, there is a potential role for automated systems or curated resources to help guide this clinical decision making.

Reference material text, such as the disease workup pages found on the Medscape reference website (http://emedicine.medscape.com/), provide clinicians with an overview of the recommended diagnostic course for a disease. Although these resources can be invaluable for clinicians in training or for casual reference, a higher-level summary of which procedures are appropriate for the diagnosis of a given disease would be beneficial for quick reference or for machine interpretation. Unfortunately, diagnostic recommendation involves more than simply declaring that a given procedure is appropriate in all cases, leading many existing resources to either give extensive detail or summarize generally over broad disease categories.

In practice, diagnostic recommendation is about understanding not just \textit{if} a procedure is a necessary part of the diagnostic process, but \textit{when}. For instance, consider the following passage describing the use of echocardiography for the diagnosis of acute coronary syndrome:

Echocardiograms play an important role in the setting of ACS. Regional wall-motion abnormalities can be identified with this modality, and echocardiograms are especially helpful \textit{if the diagnosis is questionable}. An echocardiogram can also help in \textit{defining the extent of an infarction} and in assessing \textit{overall function of the left and right ventricles}.

This passage not only gives an indication that echocardiography is an important diagnostic modality, but it also outlines specific circumstances under which it is helpful \textit{(if the diagnosis is questionable, in defining the extent of an infarction, in assessing overall function of the left and right ventricles)}. In this work we explore the automatic extraction of these conditional recommendation statements from reference material text, towards the generation of diagnostic guideline summary tables. We model diagnostic guideline generation as a series of three tasks: 1) mention-level recommendation detection, 2) conditional relation extraction, and 3) across sentence entity linking. Our results show that we can perform each task with above baseline performance, and that their output can be combined to produce concise procedure recommendation summary charts.

Related Work

While automatic extraction of conditional diagnostic guidelines – to the best of our knowledge – is a new task, its subtasks are closely related to a set of well-defined natural language processing tasks that have been applied in clinical settings, including sentiment analysis, event extraction, coreference resolution/entity linking, and text summarization.

Sentiment analysis. The recommendation identification task that we examine in this work is similar to the sentiment analysis task. Sentiment analysis, also called opinion mining, aims to determine an author’s attitude towards a specific
topic or the overall contextual polarity of a document. There has been a lot of existing work on mining reviews or other expressions of opinion on the Web\textsuperscript{1}. Particularly, in the medical informatics domain, Wallace et al.\textsuperscript{2} has conducted a large-scale sentiment analysis of online physician reviews using a statistical model guided by a small amount of annotated data. The three-way recommendation classification problem proposed in this paper is an ensemble of two typical sentiment analysis tasks: polarity classification and subjectivity identification. The former is to determine whether the expressed opinion is positive or negative (recommended vs. not recommended), whereas the latter is to determine whether a mention contains subjective information at all (recommended/not recommended vs. unspecified) and which part of the text is subjective\textsuperscript{1}. Existing approaches to sentiment analysis include those based on opinion lexicons (e.g. ontologies, semantic networks), statistical models (e.g. support vector machines, latent semantic analysis), and combinations of the two\textsuperscript{1}. In this work we present a hybrid approach to recommendation identification, where a set of recommendation keywords (lexicon) is derived from development data and used in conjunction with other word, concept, and NLP-derived features for supervised classification.

**Event extraction.** Event extraction aims to extract information about entities and the role they play in an event. The second task in our pipeline, conditional relation extraction, can be viewed as an event extraction task, where we aim to determine whether there is a conditional relation between a (recommended) procedure and a certain constraint such as a patient’s condition. There has been a substantial amount of work on event extraction published in the past decade. A typical example is molecular event extraction from biomedical literature\textsuperscript{4}. Another example in the medical informatics domain is the recent work of Botsis et al.\textsuperscript{5} on vaccine adverse event extraction from vaccine safety reports. To extract events, we need to identify the trigger/anchor word that signifies an event, and to figure out the role of each theme argument involved in the event. The presence of a candidate trigger may not necessarily suggest a relation between entities, and the problem becomes challenging when multiple potential triggers appear in the same text unit\textsuperscript{4}. Most existing approaches to event extraction decompose the task to independent classifications of events and arguments. For instance, the winner of the BioNLP’09 shared task on event extraction\textsuperscript{6} first extracts a list of candidate triggers and then determines for each pair of triggers and entities whether one is a theme of the other (with around 50% F-score).

**Coreference resolution.** The entity grouping task we explore has parallels with several related but subtly different tasks, such as coreference resolution, entity linking, and normalization. Coreference resolution determines whether two concepts are linked by an equivalence relation. For example, in the following two sentences ‘Color Doppler valve analysis during transesophageal echocardiography (TEE) can be used to …’ and ‘In patients with poor transthoracic echocardiographic images, TEE may be used to …’, the two mentions of procedures are equivalent as they refer to the same entity. A number of systems have been developed for coreference resolution in clinical texts\textsuperscript{6,7}, including those based on handcrafted rules (e.g. keywords, regular expressions, spelling correction, abbreviation expansion), supervised learning (e.g. maximum entropy classifier, support vector machines, hidden Markov model), and a mix of the two, with best-reported F-score of around 70%.

Entity linking, the task of linking each mention found in the text to a structured database such as UMLS, is done as part of UMLS concept extraction and as such is a well-studied task within the medical informatics literature\textsuperscript{10,11}. Because linking mentions back to a database implicitly groups them by the database categories, entity linking is similar to both the coreference task and the entity grouping task examined here. However, unlike these tasks, entity linking is constrained by the structure of the database, potentially leading to grouping concepts at a level inappropriate for the given task\textsuperscript{12}.

**Text summarization.** The production of conditional diagnostic guidelines across a document is one form of a text summarization task. The majority of related work in the medical informatics domain focuses on extractive summarization of electronic health records\textsuperscript{13}, where summaries are created by borrowing phrases or sentences from the original input text. In clinical summarization, normalization of words to (groups of) concepts has only recently been investigated, primarily based on well-defined ontologies\textsuperscript{14}. Our recommendation extraction pipeline built on top of clinical concept extraction sheds interesting light on identifying and aggregating similar information at a more abstract level\textsuperscript{15,16}.

**Material and Methods**

We model the production of conditional diagnostic guidelines as a pipeline of three tasks. First, each mention of a diagnostic procedure in the text is labeled as either suggesting that the procedure should be performed, suggesting that it should not be performed, or giving no recommendation one way or another. Second, we link each those mentions labeled as positive or negative recommendations with conditional statements explaining the scope of this diagnostic
recommendation. Finally, we perform a coreference task to link mentions from different sentences, producing a set of document-level recommendation statements for a given diagnostic procedure.

Data. The Medscape website provides disease-centered documents containing explanations and guidelines for the presentation, diagnosis, and treatment of these diseases. To build our dataset, we extracted the text from a set of 33 Medscape diagnostic workup pages describing cardiac ailments. Annotation was then performed to identify procedure mentions in the text and to label whether the given mention is recommending that the procedure be performed. To ease the annotation process, mentions of procedures were initially detected by simple string matching against all concepts with procedure-related semantic types (e.g., “laboratory procedure”) from the Unified Medical Language System (UMLS). Although this process provided an initial set of mentions for annotations and annotators were asked to use this as a guideline for which procedures were considered, annotators had the option to correct cases where the initial mention identification was incorrect. As a second aspect of annotation, procedure mentions were linked with conditional statements that dictated when the procedure should be undertaken. Conditional statements were represented by the head of its syntactic phrase, usually a preposition. Finally, annotators grouped together all mentions in the document that referred to the same procedure. Annotators have linguistic but not medical background. Although annotators were not medical experts, practicing clinicians were on hand to clarify any uncertainty. The brat tool was used to produce the annotations.

Figure 1 shows an example of this annotation style. The annotations in the figure combine in different ways to produce a gold standard for our three tasks. For the recommendation task, each procedure (e.g., enhanced external counterpulsation) is given a recommendation label of either unknown (no positive or negative recommendation stated in the sentence), recommended, or not recommended. For the conditional task, each trigger-procedure pair in the sentence is a candidate conditional statement, with annotators linking together true conditionals (e.g., in-EECP). No explicit example of the entity linking task is given in Figure 1. However, since the linking task compares all procedures in the document, the two procedures in Figure 1 would be checked to see if they refer to the same procedure in the given context (in this case they do not).

Overall, a total of 926 procedure mentions appeared in our dataset over a span of 2382 total sentences. Of those, 295 were recommendation statements. A total of 1103 mention-trigger pairs were present, 174 of which were marked as true conditional statements. Additionally, a small set of unannotated data from the same domain was examined for the development of features and algorithms. This data was used for development only and was not considered in training or evaluation.

Recommendation Identification. Because we wish to associate procedures with conditional recommendation statements, we perform the recommendation annotation task for each individual mention of a procedure in the text. This allows us to predict that a procedure is recommended in some instances while not recommended in others, a distinction that would be lost in document-level recommendation classification. We model the recommendation identification task as a three way (Recommended, Not Recommended, Unspecified) mention-level supervised classification task. To learn a classification model we derive features from the surrounding context in three broad categories: concept features, word features, and NLP-derived features.

Concept Features. To extract concept-based features, we first attempt to link each procedure to a UMLS concept. To do so, we use a proprietary concept extraction procedure based on identifying longest common subsequence matches between the text and candidate UMLS concepts. Although standard concept extraction procedures such as MetaMap and cTAKES exist, we chose our concept extraction procedure because internal tests suggested that it outperformed the freely available alternatives. Once concepts are identified, several features are extracted. For the concepts associated with the procedure of interest, the UMLS concept unique identifier (CUI) and its semantic and category names are retained as features. Additionally, features are produced from the CUIs of all other concepts present in the sentence. Features from concepts in the sentence were differentiated from those associated with the procedure.
concept. Finally, a Boolean feature captured whether or not any of the other concepts in the sentence were also procedures.

**Word Features.** Following previous text classification work, several feature sources were based simply on the presence of certain words in the sentence. First, a small set of both positive and negative recommendation keywords (e.g., recommend, critical, prognostic) was derived by examining a held out set of development data. One feature indicated whether any of the positive recommendation keywords were present, while another indicated the presence of negative keywords. A Boolean feature indicated whether the surrounding sentence mentioned the disease that was the theme of the document. The three words before and after the mention text were captured, as well as their relative position. Unigram and bigram features were extracted for all words in the surrounding sentence.

**NLP-derived features.** The final set of features required more linguistic interpretation. The Stanford CoreNLP toolkit (Version 3.4.1) was used to obtain a syntactic parse of the sentence and to assign part-of-speech labels. Using this information, the main verb of the sentence and any adjectives modifying the mentioned procedure were extracted as features. Positional features were extracted that indicated whether the procedure was in a list or conjunction with other procedures. Finally, the dependency parse of the sentence was used to determine if the mention was governed by a negation modifier.

**Conditional Identification.** Given a procedure mention with a positive or negative recommendation, the conditional identification task attempts to extract any part of the surrounding sentence that puts constraints on that recommendation. In order to obtain a set of candidate conditional snippets, we extract a series of trigger words that could possibly be the syntactic head of a conditional phrase. A list of possible trigger terms was generated by starting with all prepositions found in the document collection then examining development data to supplement and refine this list. Once a set of candidate triggers is identified, the task is to make a binary decision for each mention-trigger pair as to whether the phrase headed by the trigger word is a conditional statement modifying the procedure recommendation. As with recommendation identification, we model this task under a supervised learning framework.

Since many of the features used for recommendation identification potentially have discriminative value for conditional identification, the conditional identification feature vector starts by incorporating modifications of these features. This feature set is then expanded to include features capturing the relationship between the conditional phrase and the procedure and those based on the conditional phrase itself. A summary of the feature set is shown in Table 1.

**Table 1. Conditional extraction feature set.**

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Description</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Features</td>
<td>Features based on UMLS concepts extracted from the text.</td>
<td>procedure CUI, semantic name, category name; Concepts in sentence; Concepts in conditional, other procedure concepts</td>
</tr>
<tr>
<td>Word Features</td>
<td>Features based on the words in the sentence</td>
<td>keywords; disease in sentence; words near procedure; words in conditional; trigger word; unigrams; bigrams; normalized number of words in conditional</td>
</tr>
<tr>
<td>NLP-based features</td>
<td>Features based on linguistic analysis</td>
<td>verbs; modifying adjectives; positional features; trigger governed by other trigger words</td>
</tr>
<tr>
<td>Relationship features</td>
<td>Features capturing the relationship between the procedure and trigger</td>
<td>procedure is in the conditional; normalized distance from trigger to procedure; procedure and trigger have same syntactic parent</td>
</tr>
</tbody>
</table>

**Entity Grouping.** In order to turn a set of conditional recommendation statements into a recommendation chart, mentions that refer to the same conceptual entity must be linked together. While linking each mention to an entry in
an existing knowledge base such as UMLS implicitly groups them, this linking is often imperfect due to differences in word form, part of speech, or granularity between what is given in the knowledge base and what is expected in the document context. For instance, while the terms echocardiogram and echocardiography each have their own UMLS concept, they will often be used interchangeably in procedure recommendation statements. Similarly, if the term echocardiography appears after the term transthoracic echocardiography in the text they may be used to refer to the same procedure, even though normally one is a subset of the other. Given this, linking to UMLS is an insufficient method for our entity grouping task, which could be seen as more similar to traditional coreference resolution. However, unlike most coreference tasks that consider pronominal instances, in this work we are only concerned with grouping explicit (non-pronominal) mentions in the text.

We developed a new algorithm that iteratively groups mentions of procedures into clusters based on their string similarity. Given two mentions of procedures: $S = <s_1, s_2, ..., s_M>$ of $M$ words, and $T = <t_1, t_2, ..., t_N>$ of $N$ words, we define their longest common prefix as $LCF(S, T) = <p_1, p_2, ..., p_L>$, where $L$ is the largest subset of words from $S$ that found a partial match in $T$, and $p_i$ is a partial match of a word $s_i \in S$ to a word in $T$. A word $s_i$ in $S$ is said to partially match a word $t_j$ in $T$, if the length of their longest common prefix $p_i$ is above the threshold: $|p_i| \geq \tau$. When $\tau = 1.0$, this reduces to finding exact matches to words of $S$. We used relatively large $\tau$ (e.g. 0.9) for recognizing words with different forms but sharing the same root and meaning. Similar to the longest common subsequence matching problem, LCF can be computed using dynamic programming in quadratic time relative to the length of the sequences. Two procedure mentions are linked together if their LCF equals either of the two, making it possible to group coarse- and fine-grained concepts such as echocardiography and transthoracic echocardiography under the same category.

Results

Although we examine the performance of each task separately, the training and evaluation of the recommendation and conditional identification tasks follow the same basic framework. In each case, the LIBSVM package (Version 3.17) was used to train a Support Vector Machine (SVM) classifier using a linear kernel. Other kernel types were examined as well, with no impact on performance. Performance was then assessed via leave-one-out cross validation. Conversely, since our entity grouping method is based on predefined rules, no model training was required, and evaluation was done directly on the entire dataset.

Table 2. Results of the recommendation identification task.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.649</td>
<td>0.575</td>
<td>0.610</td>
</tr>
<tr>
<td>Our system</td>
<td>0.718</td>
<td>0.675</td>
<td><strong>0.696</strong></td>
</tr>
<tr>
<td>Minus Word-based</td>
<td>0.559</td>
<td>0.585</td>
<td>0.571</td>
</tr>
<tr>
<td>Minus Concept-based</td>
<td>0.690</td>
<td>0.642</td>
<td>0.665</td>
</tr>
<tr>
<td>Minus NLP-based</td>
<td>0.713</td>
<td>0.656</td>
<td>0.683</td>
</tr>
</tbody>
</table>

The results for the recommendation identification task are given in Table 2. As to our knowledge no other systems for mention-level procedure recommendation exist, we have no existing baselines for comparison. We thus compare to a bag-of-words baseline where only the unigrams and bigrams in the sentence are considered as features. As shown in the table, the proposed system is able to significantly outperform the baseline when using the entire feature set. To get a better understanding of the contribution of each features source, a feature ablation study was also conducted. As shown, removing any of the feature sources causes a loss in performance, suggesting that each feature type played a role in the prediction. However, removing NLP-based or concept-based features caused relatively modest reductions in performance. Conversely, the word-based features were shown to be most critical, as without them the classifier yielded worse than baseline performance.

For the conditional identification task, we present two views of the performance. To understand how well the classifier performs the task in a vacuum, we present the results of the task (and an associated baseline) when ground truth recommendation labels are known. However, since in a true usage scenario it is necessary to perform the recommendation identification task first, we also present overall performance numbers for a pipelined system that
performs conditional identification on the output of the recommendation identification task. Table 3 gives the results of the conditional recommendation task. Because a bag-of-words baseline would not give the classifier any information relating the procedure to the conditional, we instead chose a baseline trained on all of the word-based features given in Table 1, giving it a wider range of information. In the case where the ground truth labels are known, the learned system significantly outperforms the baseline. Similarly, the pipelined approach also outperforms the baseline, even though performing the tasks in sequence propagates the errors of the recommendation task and significantly hurts performance.

Table 3. Results of the conditional identification task.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.509</td>
<td>0.477</td>
<td>0.493</td>
</tr>
<tr>
<td>Conditional</td>
<td>0.709</td>
<td>0.644</td>
<td><strong>0.675</strong></td>
</tr>
<tr>
<td>Identification Only</td>
<td>0.581</td>
<td>0.518</td>
<td>0.547</td>
</tr>
</tbody>
</table>

The goal of the entity grouping task is to assign each mention to one of several non-overlapping sets, making evaluation less straightforward. This is a problem common to general coreference resolution approaches, where several evaluation metrics have been proposed. In this work we adopt one of the more frequently applied evaluation metrics, b-cubed score\(^{24}\). B-cubed score calculates a local precision and recall score for each mention based on the overlap between its predicted set and the ground truth set, and then produces final precision and recall numbers by aggregating over all mentions.

Table 4 shows the results of the entity grouping task. We report only a single result for our system, as the entity grouping task in not dependent on the output of the other two tasks, leading results to be the same in both the pipelined and non-pipelined scenarios. As shown, overall entity grouping performance was good, with F-measure of 0.938. We present two baselines as a means of comparison. The first baseline presents performance when grouping was done by simple string match. To get a sense of the performance of UMLS entity linking, we ran our concept extraction procedure to perform entity linking as the second baseline. Unsurprisingly, the baseline methods performed well on instances in which the terms were the same or quite similar, but less well in cases of linguistic variation. In both cases, the baseline method produced very high precision but poor recall, and produced overall F-measure significantly worse than the proposed method.

Table 4. Results of the entity grouping task.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>String Match Baseline</td>
<td>0.998</td>
<td>0.621</td>
<td>0.766</td>
</tr>
<tr>
<td>Entity Linking Baseline</td>
<td>0.990</td>
<td>0.667</td>
<td>0.798</td>
</tr>
<tr>
<td>Our System</td>
<td>0.924</td>
<td>0.953</td>
<td><strong>0.938</strong></td>
</tr>
</tbody>
</table>

Discussion

In this work we examined the recommendation identification, conditional linking, and entity grouping tasks separately, with the intent that they be combined into a larger diagnostic summary generation pipeline. This is consistent with many other multitask extraction scenarios, where pipelined approaches have been shown to be effective\(^{25}\). However, this pipelined approach potentially ignores interdependencies between the tasks that may be exploited to improve performance. In particular, it is clear that the conditional identification task is dependent on the output of the recommendation identification task, but it might also be the case that recommendations could be better identified if the presence of relevant conditional statements were detected. To explore this option, we attempted a pilot study in which the two tasks were combined with a Markov Logic Network\(^{26,27}\) based approach. Although we do not go into details about this approach here, ultimately the results indicated that the joint model produced results comparable but not superior to the presented pipelined approach. While this attempt at tighter task integration was unsuccessful, future work may wish to further examine methods that tie the tasks together.
Once each task has been performed, a diagnostic recommendation chart can be produced from the output. Several minor heuristics were applied to decide on the layout and presentation of the charts. The name used to label each procedure was selected from one of the ways in which it appeared in the text, with the most general name that mapped to a UMLS procedure given precedence. Procedures were ordered on the chart based on the order of their first appearance in the document. Conditional statements were similarly sorted by their original document ordering. To extract the conditional statement itself, the syntactic parse of the sentence was used to identify the largest phrase headed by the trigger term. Finally, a simple heuristic was used to determine if a procedure was generally (unconditionally) recommended. If at least two recommendation statements appeared for the procedure that were not linked to a conditional and none of these statements disagreed (i.e., recommended vs. not recommended), the procedure was marked generally recommended or generally not recommended, as applicable.

**Acute Coronary Syndrome Procedure Recommendation Chart**

**Electrocardiography (ECG)**
- Generally recommended.
- Recommended for angina.

**Chest radiography**
- Recommended in assessing cardiomegaly and pulmonary edema.

**Echocardiography**
- Generally recommended.
- Recommended in the setting of ACS.
- Recommended in defining the extent of an infarction and in assessing overall function of the left and right ventricles.
- Recommended in assessing overall function of the left and right ventricles.
- Recommended if the diagnosis is questionable.
- Not recommended in patients whose symptoms have resolved.

**Myocardial Perfusion Imaging**
- Recommended in other testing modalities.

**Angiography**
- Recommended in Myocardial Infarction.

**Coronary Angioplasty**
- Recommended for assessing obstructive CAD in patients without diabetes.

**Cardiac Catheterization**
- Recommended in defining coronary anatomy and the extent of a patient's disease.

**Figure 2.** Example automatically generated diagnostic recommendation chart. Incorrect extractions are given in gray.

Figure 2 shows an example of an automatically generated summary. As shown, in many instances the chart is able to present a concise statement about the conditions under which each procedure should be performed. While the figure illustrates that the tasks undertaken here have the potential to produce relatively clean recommendation tables, it also highlights some of the potential sources for error. First and most critically, because of the nature of the automatic generation process several of the recommendations and conditionals are incorrect. For instance, the table contains the nonsensical statement that myocardial perfusion imaging is “recommended in other testing modalities”. Because the cost of giving an incorrect recommendation is potentially severe, automatically generated summaries would need to be accompanied by human oversight in a true human usage scenario, but may still have beneficial use cases for machine interpretation. A less severe error is the extraction of statements that, while not quite incorrect, provide no extra conditional information. For instance, Figure 2 states that echocardiography is “recommended in the setting of ACS”, which gives no further conditional information in this context. Finally, we note that some additional errors arise from parsing errors, where the extracted conditional contains more or less information than the ground truth.

To get a rough understanding of the overall quality of the automatically generated summaries, we performed a small qualitative study. For each of the 33 diseases we randomly generated a recommendation chart using either the baseline, system, or gold standard methods. The generated charts were then shown to a medical resident, who examined their correctness and coherence and assigned a rating to each one on a 1 (best) to 4 scale. Results of this study suggested that the quality of the system generated charts was at least comparable to those generated by the gold
standard, with average ratings of 1.72 and 1.50, respectively. However, because of the small scale of the study, we caution that while these results suggest that the automatically generated summaries have potential, a more robust analysis would be required to fully understand the overall quality.

While the tasks undertaken in this work can be used collaboratively to produce diagnostic guideline summaries, they each have the potential to play other roles in medical text extraction. Recommendation classification can be used separately to identify procedure recommendations at a high-level, and has overlap with other information extraction tasks of interest, such as the extraction of diagnoses and findings from the electronic health record\textsuperscript{28,29}. Similarly, conditional extraction has parallels with other relation extraction tasks, such as biological event extraction\textsuperscript{30–32}, semantic relation labeling\textsuperscript{33,34}, and the identification of treatment relations\textsuperscript{35}. Finally, coreference resolution and entity linking are frequently applied to extract and link clinical terms from various sources of unstructured text, although approaches for reference text may be different than those made on clinical notes\textsuperscript{36}.

**Conclusion**

Understanding the scenarios in which a diagnostic procedure should be performed is a difficult and nuanced task, and one in which there is a potential role for NLP-based automation and guidance. In this work we examined the extraction of conditional recommendation statements, as a first look at diagnostic recommendation summarization. This was done by examining the assignment of mention-level recommendation labels, linking these mentions with conditional recommendation statements, and grouping them based on their underlying referent. Our results show that each of these tasks can be performed with above baseline accuracy, and the resulting output can be combined into concise summaries of the diagnostic procedure. While our automatically generated summaries are likely to contain more noise than those generated from gold standard human annotations, they represent a first step toward the automated extraction and understanding of the diagnostic workflow.

While the conditional summaries produced here attempt to give an overview of when procedures are appropriate, they do not impose a relative order on when each procedure should be performed. As such, the current recommendation summaries still require that some inference be performed by the clinician to choose the most appropriate test for a given scenario. While this does not necessarily diminish their usefulness to human readers, it can be a potentially problematic omission for machine interpretation. Given this, future work will focus on the expansion of this summarization framework towards a fully formed flowchart of diagnostic procedures that attempts to give a conditional roadmap to the entire diagnostic process.

**References**


Using Monte Carlo/Gaussian Based Small Area Estimates to Predict Where Medicaid Patients Reside

1Jess J. Behrens, M.Sc., 2Xuejin Wen, Ph.D, 1Satyender Goel, Ph.D., MBA, 2Jing Zhou, Ph.D, 2Lina Fu, Ph.D, 1Abel N. Kho, MD, MS

1Center for Health Information Partnerships, Northwestern University, Chicago, Illinois, 2PARC, A Xerox Company, Rochester, New York

Abstract

Electronic Health Records (EHR) are rapidly becoming accepted as tools for planning and population health1,2. With the national dialogue around Medicaid expansion12, the role of EHR data has become even more important. For their potential to be fully realized and contribute to these discussions, techniques for creating accurate small area estimates is vital. As such, we examined the efficacy of developing small area estimates for Medicaid patients in two locations, Albuquerque and Chicago, by using a Monte Carlo/Gaussian technique that has worked in accurately locating registered voters in North Carolina11. The Albuquerque data, which includes patient address, will first be used to assess the accuracy of the methodology. Subsequently, it will be combined with the EHR data from Chicago to develop a regression that predicts Medicaid patients by US Block Group. We seek to create a tool that is effective in translating EHR data’s potential for population health studies.

Introduction

Electronic Health Records (EHR) are a promising data source for examining population health and for community health needs assessments1,2. For high density populations, zip codes may be used as the units for location analyses, but zip codes are widely considered to be insufficiently granular for modelling environmental/human interactions3,4,5,6,7. While geo-statistical methods exist for interpolating probable location from known points8,9,10, the literature is sparse on evaluating how accurate these techniques really are at predicting where the modeled event occurs. Understanding the accuracy of these methods, at spatial & demographic resolutions that are meaningful to health related processes, is vital for epidemiological studies based on EHRs to be successful.

We set out to evaluate the accuracy of one such technique for predicting probable patient location, a Gaussian Geo-statistical & Monte Carlo methodology that has proven effective for estimating probable voter location11. For this analysis, we selected Medicaid status as our condition of interest. We selected Medicaid status for two reasons. First, recent expansion of Medicaid status to new populations presents a controversial effect of the Affordable Care Act ripe for analyses of effects on population health12. Secondly, it represents a definite, unique indicator of patient socio-economic status that is most likely also associated with both patient health outcomes & exposure to potential health related environmental influences13,14. Thus, the goal of our study is to ensure that patient location can be accurately imputed from zip code aggregated EHR data using U.S. Block Group Census counts as a tool to weight that imputation, just as it was for voter location.

Methods

Using only registered Medicaid patients in 2 different cities (Chicago & Albuquerque), we developed small area estimates of Medicaid patients for both study areas from aggregated zip code patient counts to block group using a combination Monte Carlo/Gaussian Geo-statistical simulation technique. Chicago Medicaid patients were represented using HealthLNK EHR records. HealthLNK represents a total of 6 years (2006-2011) of de-identified & de-duplicated Electronic Health Records (EHR) obtained from 6 different sites across Chicago and are thus only a sample of the total Medicaid patients in the Chicago area16. The geo-imputation for Chicago was meant as a comparison and will be used in future steps of the project. Conversely, in Albuquerque, where we have all Medicaid records & patient address data, we compared accuracy to another study done using the same methods but on registered voters in central North Carolina. The accuracy assessment in Albuquerque will be done separately over two years (2012 & 2014), using the most recent address for each Medicaid patient in each year to represent where that patient lives.
We address matched Albuquerque Medicaid patients using ArcGIS 10.3 and subsequently aggregated by zip code. Because the methodology weights probable patient location using U.S. Census Block Group counts, Albuquerque zip code to block group geographic coincidence was established in ArcGIS using a spatial join. We imputed probable patient block group location by performing a Monte Carlo simulation that uses limited personal data (age, gender, & ethnicity) & associated US Census Block Group totals to establish the probable average number of zip code aggregated Medicaid patients that live within each associated block group. These probable Medicaid patient block group averages were distributed among associated census blocks proportionally & kriged in ArcGIS. A krig is a raster based statistical surface, similar to a digital elevation model, where the raster cells represent a probability, in this case the number of Medicaid patients living there. The resulting krig was fed into a Gaussian Geo-statistical Simulation to generate an average & standard deviation probability raster to evaluate the accuracy of the predicted average number of Medicaid patients living in each raster cell (Figure 1).

We assessed accuracy for each of the three years, separately, using the Root Mean Square Error (RMSE), Error Product, and Error Product/RMSE for each (Table 1). We compared the values for each year to the North Carolina Voter results from our prior study. RMSE is a common measure of accuracy which is calculated as the square root of the average squared error for each prediction made. Since the results here involve geography, the RMSE assesses the average number of patients that the raster has over or under predicted at every point/raster cell within the study area. It follows logically that the smaller the RMSE the better. The Error Product measures how consistent the results are. It has two separate components, which are multiplied together. The first is the percentage of checked locations used in the RMSE that fall within 3 Standard Deviations of the mean. The second is the percentage of checked raster cells with a predicted average number of Medicaid patients that is greater than the 1 standard deviation measured at that same location. The goal is to have 100% of observations comply with each of these criteria, a situation that would yield an Error Product of 1. These two measures are important because they are an indicator of the quality of the simulation as an approximation of the Medicaid patient distribution. Deviations that are far from 1 indicate that the methodology is faulty, and that the RMSE should not be trusted regardless of its magnitude. Finally, the Error Product RMSE is simply the Error Product divided by the RMSE. Since the goal for each measure is a value of 1, where an RMSE of 1 would indicate that the raster has an error of at maximum one Medicaid patient at any given point in the study area, the Error Product RMSE should also be evaluated relative to a value of one.

Results

A total of 108,308 Albuquerque Medicaid patients who had received care within the last year and who were living within 482 block groups were selected for the study. This included, by year, 27,143 Medicaid patients in 2012 & 81,165 Medicaid patients in 2014. In Chicago, 88,198 Medicaid patients were selected who fall within 217 zip codes in the Cook, DuPage, & Will County areas. These patients represent all patients in those zip codes whose final insurance status in HealthLNK was Medicaid.
Table 1 shows the results of the accuracy comparison for Albuquerque & the North Carolina Voter dataset project.

**Table 1.** RMSE, Error Product, & Error Product RMSE, Albuquerque Medicaid Patients & North Carolina Voters

<table>
<thead>
<tr>
<th>Accuracy Measure</th>
<th>Albuquerque Medicaid</th>
<th>North Carolina – 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2012</td>
<td>2014</td>
</tr>
<tr>
<td>RMSE</td>
<td>5.3</td>
<td>11.61</td>
</tr>
<tr>
<td>Error Product</td>
<td>0.853</td>
<td>0.859</td>
</tr>
<tr>
<td>Error Product RMSE</td>
<td>0.161</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Given that the number of Albuquerque Medicaid patients in each year number between around 30 to 80,000, and are represented in a Gaussian raster with a resolution less than 0.01 mi², the most appropriate comparison to the North Carolina Voter project is at both 10% & 30% in the Urban Plus study area, which has 27,125 & 81,367 total voters, respectively. As table 1 shows, the RMSE is virtually identical in both study areas, with greater variability in error at 10% than 30%, which may be caused by sampling error at these lower population levels.

The higher RMSE in Albuquerque in 2012 may also be a result of population distribution. Chicago is very densely populated near the lake, & that population drops as one moves west, toward the suburbs. While

---

**Figure 1.** Probable Average Medicaid Patients, Albuquerque & Chicago Study Areas
Albuquerque has areas that are similar in density to Chicago, these areas are much smaller. Furthermore, the population density drops at a much more rapid rate as one moves out from these areas and into the surrounding desert. These stark contrast between the population distribution in Chicago & Albuquerque can be seen in a block group map of the two study areas. As block groups increase in size, the population density decreases. The krig, which is used as a base for the Gaussian simulation, can account for much of these variations, but it may be that the stark contrast in population distribution around Albuquerque may stretch its limits.

The reason for this stems from the algorithm and mathematical methods that constitute a krig. In effect, it is the same thing as fitting a curve in traditional statistics, it’s just being applied to data that has the added complexity of being spatially explicit. As is common in non-spatial statistics, binning, or grouping, data can create a sort of bias. As such, in a normal statistical project that requires binning, much time and attention is paid to how the underlying data is broken out into its groups. This, of course, includes the problematic question of how many bins to include. With spatial data such as that found in the US Census & ACS data, the binning choices have already been made. Because those bins also have the added complexity of representing an area on the ground with a specific extent and a specific spatial relationship to all of the other bins, the bias introduced by binning is not only found in the underlying counts, but in where the boundaries for that bin have been set. In spatial analysis, the two primary types of error introduced by binning are commonly called spatial clustering and spatial trend. Kriging corrects as much as possible for these two types of bias. By removing this type of bias, we can use the Gaussian to examine the degree to which the underlying data is actually spatially clustered minus that bias. Other forms of small area estimates don’t account for these types of error and bias, making the Gaussian methodology presented here unique.

Thus, after adjusting the observations for these two data aggregation errors, the krig uses a type of machine logic to ‘fit’ a three dimensional surface to those observations. The more normally distributed, or spatially disperse, the underlying data is, the better the equation developed for the krig will fit it. That goodness of fit will lead to a lower RMSE after the Gaussian simulation is run. As the Gaussian RMSE increases, within the confines of a similar or unchanging Error Product, that increase will most likely be due to non-normally distributed, or spatially clustered, underlying data.

Discussion

In this project, we created small area estimates of Medicaid patients using the same methods applied in two distinct geographies. When visualized on a map, our estimates correlate with known areas of low socioeconomic status (SES) in both cities. Compared with a prior validation study applied to voter registration records in North Carolina, our estimates of Medicaid patient distribution generated a slightly larger RMSE at low population counts, but essentially the same. When one considers that, as a rule, population rarely distributes itself ‘normally’ in space, the fact that Medicaid status is dependent on SES & registering to vote is not indicates that it is most likely these socio-economic factors that are responsible for the additional RMSE at low population levels in 2012. Fortunately, the RMSE increase is not that much relative to the total population being simulated & the Error Product for both projects is almost identical. These facts, taken together, strongly indicate that the krig & subsequent Gaussian simulation provide a strong model for Medicaid patient location. Furthermore, the increase in RMSE, most likely due to the clustering effects of low SES, indicates that the block group aggregate average Medicaid patients will serve as a strong dependent variable for future work by our group to apply regression analysis to estimate Medicaid patient population in areas based on socio-economic factors alone.

Our group is currently studying the impact of Medicaid expansion on diabetes outcomes across ten states, half of which are non-Medicaid expansion states. Developing accurate methods to estimate Medicaid patients per block group nationally will help us identify a comparison population of patients in non-Medicaid expansion states who would have qualified for Medicaid had their states chosen to participate in the Medicaid expansion program. Accurate imputation methods enable researchers to study the impact of policies or other external shocks on clinical outcomes when data are sparse or missing.

References

Recognizing Question Entailment for Medical Question Answering

Asma Ben Abacha, PhD & Dina Demner-Fushman, MD, PhD
U.S. National Library of Medicine, Bethesda, MD.

Abstract

With the increasing heterogeneity and specialization of medical texts, automated question answering is becoming more and more challenging. In this context, answering a given medical question by retrieving similar questions that are already answered by human experts seems to be a promising solution. In this paper, we propose a new approach for the detection of similar questions based on Recognizing Question Entailment (RQE). In particular, we consider Frequently Asked Question (FAQs) as a valuable and widespread source of information. Our final goal is to automatically provide an existing answer if FAQ similar to a consumer health question exists. We evaluate our approach using consumer health questions received by the National Library of Medicine and FAQs collected from NIH websites. Our first results are promising and suggest the feasibility of our approach as a valuable complement to classic question answering approaches.

1 Introduction

Consumer health queries are increasing at a high rate on the World Wide Web. Recent studies from the Pew Research Center show that 59% of U.S. adults searched for health information online in 2013. In the same study, we find that 35% of U.S. adults attempted to figure out what medical condition they or someone else might have by searching online resources. Consumer health questions cover a wide range of health-related topics asked generally by non-expert persons, patients or professionals. Typical information sought by consumers can be, for instance, information about a disease (e.g. “My daughter was recently diagnosed with a CTNNB1 mutation and there are only 4 published cases. Can you tell me where I may be able to find more information”), treatments (e.g. “Do you have information on the treatment of Parkinson’s disease with the use of amino acids”) or more specific questions about drugs (e.g. “is it safe to take diclofenac when taking lisinopril or aleve or extrastrength Tylenol?”).

Many websites offer online doctor consultation services in virtual hospitals (e.g. icliniccare.com, www.icliniq.com, www.jeevom.com, www.evaidya.com, www.doctorvista.com) and propose to answer medical questions, give advice or second opinions from thousands of doctors and specialists. These services attract more and more internet users seeking easy and quick answers, as well as free consultations and privacy. Many of these resources publish freely the submitted questions and doctors’ answers. New health-dedicated forums and websites continue emerging each year; and major question-answering websites such as Quora or Yahoo! Answers include important and growing health sections.

With this multitude of information, duplicate questions are becoming more frequent, and finding the most appropriate answers becomes problematic. This issue is important for question-answering platforms as it complicates the retrieval of all information relevant to the same topic, particularly when questions similar in essence are expressed differently. Finding similar questions is also important for the users as is implied by the 2013 survey, in which 16% of the users tried to find others with the same health concerns.

Efficient automatic approaches are therefore required to detect similar questions both at search time and at question submission time. In this paper we propose a new approach for the detection of similar questions based on Recognizing Textual Entailment (RTE). RTE is an important component in Natural Language Understanding and in Question Answering. Harabagiu and Hickl argue that RTE can enable question answering systems to identify correct answers with greater precision than keyword or pattern based methods [1].

We particularly tackle the detection of a most similar Frequently Asked Question (FAQ) to a given consumer question. FAQs are both a valuable and widespread source of information. Many trusted sources, such as NIH institutes and

---

1 http://www.pewinternet.org/2013/01/15/health-online-2013/
2 Example questions received by the U.S. National Library of Medicine.
3 www.quora.com As of March 1 2016, 7,176,197 questions were posted to Quora.
4 answers.yahoo.com
centers, provide answers to frequently asked questions organized by topics. NIH provides FAQs for many health related problems such as Rare Diseases (Frequently Asked Questions About Rare Diseases\(^5\), or Alzheimer\(^6\)).

To test our approach in a use case, we obtained consumer health questions received by the National Library of Medicine (NLM). NLM receives hundreds of requests per day, e.g., from October 2013 to September 2014, NLM received a total of 102,622 requests, including many consumer health questions. Our final goal is to automatically provide an existing answer if an FAQ similar to a consumer health question received by NLM exists. If we can reliably identify similar questions, retrieving the corresponding answers is relatively straightforward. For example, this approach to answering health related questions is taken in the SimQ system [2].

In this paper, RTE is applied to find a frequently asked question similar to a consumer health questions in order to answer consumer health questions with the answers given to similar FAQs. As far as we know, RTE from medical questions has not been studied before for Question Answering. Our contributions can therefore be summarized in three points:

- We address and define the problem of RTE in medical questions for Question Answering.
- We construct automatically an RTE training corpus of medical questions and we study the impact of varying the size and nature of the training examples.
- We use different features for the RTE task in medical questions including classical similarity measures and semantic features related to the medical domain.

The remainder of the paper is organized as follows: Section 2 presents the background. Section 3 presents our approach for RTE in medical questions developed to retrieve existing FAQs and answers. In section 4, we present our methods to construct automatically an RTE training corpus, and to construct semi-automatically an RTE test corpus for medical questions. Our experiments and results are detailed and discussed in sections 5 and 6.

2 Background

Recognizing Textual Entailment (RTE) has been addressed by numerous works in the literature and in the framework of the PASCAL challenge \(^7\). Dagan et al. [3] present the RTE task and give an overview of the research efforts in this area. A detailed survey of RTE approaches is also presented in [4].

Several machine learning (ML) methods have been explored for RTE using different features, e.g., similarity measures [5, 6, 7]. Some efforts focused on the training corpora as they are an important factor for an efficient supervised learning system. Zanzotto and Pennacchiotti [8] proposed a method to expand the existing textual entailment corpora. They extracted from Wikipedia a large set of textual entailment pairs and used a semi-supervised machine learning method to make the extracted dataset homogeneous with the existing corpora. Other efforts tackled automatic generation of training corpora for a specific language. For instance, Marzelou et al. [9] proposed a method to create a Greek Textual Entailment Corpus that can be exploited for training or evaluating a system for RTE from Greek texts.

While textual entailment in open-domain has been extensively addressed in the literature, RTE has been less addressed for more restricted and specialized fields such as the medical domain. Adler et al. [10] presented a text exploration system, in which search results in the health-care domain can be navigated at propositional level according to textual entailment relation. Ben Abacha et al. [11] proposed a supervised learning method to RTE from medical texts. For the same purpose, they also proposed an automatic method for construction of training corpora from MEDLINE abstracts and compared the results obtained from the open-domain model, derived from the PASCAL training corpus, and the medical-domain model, derived from the automatically-constructed corpus. Their experiments showed the benefits of domain-related models and automatic corpus construction.

\(^5\)http://www.genome.gov/27531963 National Human Genome Research Institute, NIH.  
\(^6\)http://nihseniorhealth.gov/alzheimersdisease/faq/faqlist.html  
\(^7\)http://www.nist.gov/tac/2011/RTE/
One of the earliest question answering systems based on finding similar questions and re-using the existing answers was FAQ FINDER [12]. Some of the underlying assumptions about FAQs were that the information needed to determine the relevance of a (Question/Answer) QA pair can be found within the QA pair and that the question half of the QA pair is the most relevant for determining the match to a users question. Not surprisingly, much of the subsequent work was dedicated to question matching. Jeon et al.[13] showed that a retrieval model based on translation probabilities learned from a question and answer archive can recognize semantically similar questions. Duan et al. [14] proposed a language modeling using question topic and question focus for question search. An alternative approach, is to return a ranked list of QA pairs in response to a user’s question, treating finding an answer as a fielded search task, where the user’s question is treated as a query, and the items to be returned are QA pairs [15].

More recent research efforts have focused on retrieving similar consumer health questions. Wang et al. [16] developed a syntactic tree matching method to find similar questions for 0.5 million QA pairs from the Healthcare domain in Yahoo! Answers. SimQ [2] aims to retrieve similar web-based consumer health questions. The system uses syntactic and semantic features and reaches a precision of 72.2% and a recall of 78.0%. SimQ is used to complement the existing Q&A services of Netwellness and allows reducing response delay by instantly providing closely related questions and answers.

3 Recognizing Question Entailment (RQE)

A. Problem Definition

Textual Entailment (TE) is a directional relation between two text snippets called text (T) and hypothesis (H), expressing the fact that the meaning of T is contained in the meaning of H [3]. In a similar definition, the first PASCAL Recognizing Textual Entailment Challenge (2004-2005) defined the task of Recognizing Textual Entailment (RTE) as deciding, given two text fragments, whether or not the meaning of one text (H) can be inferred (entailed) from the other one (T) [17].

In our case, the (T, H) pairs refer to pairs of questions (Q1, Q2). Our goal in to retrieve answers to Q1 by retrieving entailed questions Q2 that have associated answers. Groenendijk and Stokhof (1984) define an entailment relation between two questions Q1, Q2 if every proposition giving an answer to Q1 is also giving an answer to Q2 [18]. Roberts (1996) called the first question Q1 as a superquestion and Q2 a subquestion (if we answer enough subquestions, we have the answer to the superquestion) [19]. In accordance with these definitions, we define question entailment as follows:

Question Entailment. Question A entails Question B if every answer to B is also a correct answer to A exactly (cf. Example 1) or partially (cf. Example 2).

Consumer health questions often contain heterogeneous information, e.g. information about the patient’s history. Our rule for ignoring additional information in the question is based on the assumption that it is not required to retrieve a correct answer (cf. Example 1, in which the fact that the patient has only central vision will not help retrieving information related to treatment of the disease). Here are two examples from our datasets:

- Example 1:
  - A1 (consumer health question): Hi I have retinitis pigmentosa for 3years. Im suffering from this disease. Please intoduce me any way to treat mg eyes such as stem cell ....I am 25 years old and I have only central vision. Please help me. Thank you
  - B1 (FAQ): Are there treatments for RP?
  - A1 → B1

- Example 2:

---

8http://netwellness.org/
9http://pascallin.ecs.soton.ac.uk/Challenges/RTE/
• A2 (consumer health question): Can sepsis be prevented? Can someone get this from a hospital?
• B2 (FAQ): Who gets sepsis?
• A2 → B2

A2 includes 2 questions about prevention and susceptibility. An answer to B2 (about the susceptibility to Sepsis) is considered as a partially correct answer to A2. In this case, we consider that A2 implies B2.

B. Learning Method for RQE

We propose a supervised machine learning approach to determine whether or not a question Q2 can be inferred from a question Q1. In order to extract relevant features, we first remove stop words and perform word stemming using the Porter algorithm [20] for all (Q1,Q2) training pairs.

**Lexical Features:**

We compute different similarity measures between the pre-processed questions and use their values as features:

- **Word Overlap:** we compute the word overlap as the proportion of words that appear in both Q1 and Q2 and normalize by the length of Q1.
- **Bigram:** we compute the bigram similarity between Q1 and Q2 as the total number of matched bigrams in (Q1,Q2) pair normalized by the number of Q1 bigrams.
- **Best similarity value:** the maximum similarity between five similarity measures: Levenshtein, Bigram, Jaccard, Cosine and Word Overlap.

**Semantic Features:**

**Negation:** we use NegEx [21] for identifying negation scope in Q1 and Q2.

**Medical entities:** we annotate all (Q1,Q2) pairs with medical entities using two supervised systems. The first system uses a CRF classifier trained on the i2b2 corpus [22] to recognize medical entities of 3 types: Problem, Treatment and Test [23]. These three medical categories are the most frequent and important categories in consumer questions. The second system is a meta-classifier [24] trained on four corpora: two clinical texts corpora i2b2 and SemEval [25] and two scientific abstracts corpora NCBI [26] and Berkeley [27] to recognize medical problems. In many cases, the focus of the consumer health question and the FAQ is a medical problem (e.g. questions about treatment, symptoms or medical exam). Starting from the obtained annotations, we generate the following semantic features for each question pair (Q1,Q2):

- Number of medical entities in Q1. Number of medical entities in Q2.
- Number of medical problems in Q1. Number of medical problems in Q2
- Number of medical entities in common between Q1 and Q2.
- Number of medical problems in common between Q1 and Q2.
- **Common medical problem:** binary feature indicating whether or not at least one medical problem from Q1 is mentioned in Q2.

4 Data

A. Automatic Construction of Training Data

We used the NLM collection of 4,655 clinical questions asked by family doctors [28] to construct our training corpus for RQE. An extract from this collection is presented below (Clinical question with NLM ID: NQ003094):
6 and 1/2-year-old girl. What’s causing her ear pain? She has a normal ear on exam and normal tympanogram. Is it just eustachian tube dysfunction? 

What is causing the ear pain in this child with a normal ear exam?

What would cause ear pain in a child with a normal exam?

History

Earache

Diagnosis

To obtain positive QE examples, we use the original form and the short form of the question as expressed in the collection.

original question → short form (4,655 positive examples)

To obtain the final corpus, we studied five different construction methods by varying the number and type of negative pairs, and we analyzed the impact of each construction method on the results. Each of these methods led to different training sets, described below:

Training set 1: 8,588 training pairs, containing 54.2% positive pairs. The remaining pairs (3,933) are negative examples collected by associating a random short form having at least one common keyword and at least one different keyword for each original question.

Training set 2: 17,898 training pairs, containing 26% positive pairs. The remaining pairs (13,243) are negative examples constructed as follows:

- original question → short form of random question having at least 1 different keyword and at least 1 common keyword
- original question → short form of random question having at least 1 different keyword and at least 1 common content
- original question → short form of random question having at least 1 different keyword

Training set 3: 13,918 training pairs containing 33.44% positive pairs. Two types of negative examples are considered in this set:

- original question → short form of random question having (i) at least one different keyword and at least one common keyword or (ii) at least one different <content> tag and one equal <content> tag (cf. example question above)
- original question → short form of random question having (i) at least one different keyword and at least one equal <content> tag or (ii) at least one different <content> tag and at least one equal keyword.

Training set 4: 18,573 training pairs with 25% positive pairs. Three types of negative examples (13,918 negative pairs):

- Same 2 types of negative pairs as in training set 3.
- original question → short form of random question having (at least one different CONTENT) OR (at least one different keyword)

Training set 5: 9,310 training pairs with 50% positive pairs. One type of negative examples regrouping the 2 types of Training set 3 (4,655 negative pairs):
original question \rightarrow short form of random question having (i) at least one different keyword and at least one common keyword or (ii) at least one different <content> tag and at least one equal <content> tag, or (iii) at least one different keyword and at least one equal <content> tag, or (iv) at least one different <content> tag and at least one equal keyword. Duplicated questions were removed.

The four examples below are extracted from our training corpora to better show the intuition behind these different construction procedures.

- <pair id="29" type="originalQ-shortRandQ" value="negative">
  <t> What is the cause and treatment of this old man’s stomatitis? </t>
  <h> What would cause a patient to have a low thyroxine and a high thyroid stimulating hormone when the patient is on Synthroid chronically? </h>
</pair>

- <pair id="37" type="originalQ-shortQ" value="positive">
  <t> Is melatonin good for anything? I don’t know anything about melatonin. I need to know the dose. </t>
  <h> Is melatonin good for anything? What is the dose? </h>
</pair>

- <pair id="4358" type="originalQ-shortRandQ" value="negative">
  <t> How should you work up someone who tried to give blood and has positive Hepatitis B core antibody? </t>
  <h> Why did the blood pressure go up with addition of Vasotec? (Was already on Procardia.) </h>
</pair>

- <pair id="10366" type="originalQ-shortQ" value="positive">
  <t> Young woman with acute Fifth disease (son also has it). She is not pregnant. She has increased swelling of her hands and her arms go numb and she has trouble sleeping. Wants to know what to do. </t>
  <h> What is the treatment for Fifth disease? </h>
</pair>

We evaluate the constructed training corpora in terms of Precision, Recall and F-measure in section 5.

B. Semi-automatic Construction of Test Data

For test pairs, we collected two types of test data: (i) pairs of manually validated questions from the NLM collections and (ii) pairs of questions including FAQs retrieved online with a manual search of NIH websites. We constructed the two parts of our test corpus using the following methods:

We constructed the first part using two collections of (i) 300 consumer health questions annotated with the focus of the question and (ii) 349 FAQs from NIH websites, in two steps:

1. Extracting the question pairs that have the same focus (medical problem).

Here are two examples from the first part of our test corpus:

- <pair id="6" type="Part1" value="TRUE">
  <t> No. hi my name is NAME. I’m currently working with Friends Community Center in Hollywood California and I was wondering I came across some of you healthy tip fliers for HIV/Aids treatment .at the moment we have a study going on that helps HIV positive transgender women into HIV quality care .so it would be great to have some more information on HIV/Aids treatment </t>
  <h> How is HIV/AIDS treated? </h>
</pair>

- <pair id="66" type="Part1" value="FALSE">
  <t> recovery after stroke?. what is the pattern of recovery after stroke? </t>
  <h> Is there any treatment for Stroke and Atrial Fibrillation? </h>
</pair>

For the second part, we searched online for FAQs for 116 from the remaining consumer health questions from NIH websites to construct our positive pairs. For negative pairs, we selected for each consumer health question, a random FAQ from the list of 116 found online. Here are two examples from the second part of our test corpus:

- <pair id="79" type="Part2" value="TRUE">
  <t> Help for my diagnose. I have been diagnosed with SCA3. I was wondering if MedlinePlus is able to help me with resources that I may need on my journey through this disease? If not, can you help me find an organization or association that can help me. </t>
  <h> Where can I find additional information about SCA3? </h>
</pair>

315
Our final test corpus contains 302 pairs of questions consisting of 173 negative pairs and 129 positive pairs.

5 Evaluation

In this section, we evaluate the corpus construction methods as well as several algorithms to recognize textual entailment between questions.

Table 1 presents the results in terms of Precision (P), Recall (R) and F-measure (F), obtained using different training corpora. We tested different configurations (i) by adding new types of negative pairs (e.g. training set 3 vs. training set 4) to evaluate the impact of the new example types and (ii) by regrouping types of pairs to evaluate the impact of reducing the number of negative examples (e.g. training set 3 vs. training set 5).

<table>
<thead>
<tr>
<th>Training Data</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>75.0</td>
<td>75.2</td>
<td>75.0</td>
</tr>
<tr>
<td>Set 2</td>
<td>72.5</td>
<td>71.9</td>
<td>70.7</td>
</tr>
<tr>
<td>Set 3</td>
<td>73.4</td>
<td>73.5</td>
<td>73.2</td>
</tr>
<tr>
<td>Set 4</td>
<td>71.6</td>
<td>71.2</td>
<td>70.1</td>
</tr>
<tr>
<td>Set 5</td>
<td>72.7</td>
<td>72.8</td>
<td>72.4</td>
</tr>
</tbody>
</table>

Table 1: Results of a SVM classifier trained on five datasets having different sizes and types.

Table 2 presents the obtained results using four statistical learning algorithms that are usually used for RTE (SVM, Logistic Regression, Naive Bayes and J48). The best results are obtained using the SVM classifier (75% F-measure). The Logistic Regression classifier gives a slightly lower F-measure on our dataset (74.7%).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>75.0</td>
<td>75.2</td>
<td>75.0</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>74.7</td>
<td>74.8</td>
<td>74.7</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>73.1</td>
<td>72.5</td>
<td>71.5</td>
</tr>
<tr>
<td>J48</td>
<td>70.9</td>
<td>70.2</td>
<td>70.3</td>
</tr>
</tbody>
</table>

Table 2: Results of different classifiers trained on Set1 of training pairs.

6 Discussion

By testing the impact of different automatic corpus construction methods to train RQE classifiers we found that increasing the number of negative examples does not always improve the results. The nature and size of the training corpus obtained by our best automatic construction method (8,588 question pairs with 54.2% of positive pairs) provided a good start for building a system for question entailment recognition that can be used as an important support to find additional answers that complement classical question-answering results.

While we obtained relatively good results in our evaluation, several challenges have still to be tackled to enhance further the performance of similarity-based approaches. If we consider the example of questions Q1 and Q2 below, they are not linked by an entailment relation even though they share three different keywords (cause, dry, mouth).

- **Q1 (Consumer health question):** treatment for dry mouth caused by necessary medicine. My provider can’t help (I asked.) I am intolerant of all the sugar alcohols such as maltitol, sorbitol, xylitol, etc. and need something for dry mouth caused by med which I have to take. Biotene products help for only about two minutes.
- **Q2 (FAQ):** What causes dry mouth?
• *Entailment*: false.

More generally, relying on number of shared keywords and medical entities provided a high recall upper bound for the recognition of question entailment, however, it did not reach very high precision due to the lack of advanced features such as the identification of the answer type (e.g. treatment in Q1) and the main semantic relations (e.g. *causes* in Q2).

7 Conclusion

In this paper, we described our approach for recognizing question entailment (RQE) in order to answer new questions using existing Question-Answer pairs. We presented an automatic method for the construction of training corpora for RQE and a semi-automatic method for the construction of a test corpus for medical questions. Our experiments confirm the feasibility of medical question entailment even with a small set of features. In future work, we plan to extend our test corpus and to include more semantic features. We also plan to study the adaptation of our training corpus consisting of doctors’ questions by including a small number of consumer health questions (e.g. questions asked by non experts, patients) and manually collected FAQs.

References


Hypothesis-Free Search for Connections between Birth Month and Disease Prevalence in Large, Geographically Varied Cohorts

John P. Borsi

1Explorys, an IBM Company, Cleveland, OH

Abstract

We have sought to replicate and extend the Season-wide Association Study (SeaWAS) of Boland, et al. in identifying birth month-disease associations from electronic health records (EHRs). We used methodology similar to that implemented by Boland on three geographically distinct cohorts, for a total of 11.8 million individuals derived from multiple data sources. We were able to identify eleven out of sixteen literature-supported birth month associations as compared to seven of sixteen for SeaWAS. Of the nine novel cardiovascular birth month associations discovered by SeaWAS, we were able to replicate four. None of the novel non-cardiovascular associations discovered by SeaWAS emerged as significant relations in our study. We identified thirty birth month disease associations not previously reported; of those, only six associations were validated in more than one cohort. These results suggest that differences in cohort composition and location can cause consequential variation in results of hypothesis-free searches.

Introduction

The human urge to assign importance to birth season is well-documented: astrologists have been attempting to divine human fates based on birth timing for millennia. Such endeavors were put on solid scientific footing in 1929, with the publication of a work on the connection between mental disorders such as schizophrenia and birth month. Little doubt remains that birth month does have a measurable impact on many facets of life; over 250 studies had confirmed the importance of birth season before the year 2000. Links have been established between disease prevalence and birth month for numerous conditions: allergies and rhinitis, reproductive performance, attention deficit hyperactivity disorder (ADHD), dermatitis, Crohn's disease, and otitis media, among others. Studies have also suggested connections between birth month and height, life expectancy, and life events. Potential explanations for the observed links have incorporated diverse causes such as neonatal vitamin D exposure, exposure to allergens, and the impact of social age.

In 2015, Boland et al. implemented a hypothesis-free, phenome-wide method to systematically identify associations between birth month and disease prevalence. Their work is part of a growing acceptance of using electronic health record (EHR) data to conduct retrospective studies. EHRs have been mined to better understand health care utilization of diabetic patients, detect adverse drug events, and find health care fraud. “Hypothesis-free systems,” algorithms which proceed systematically through a dataset without a priori hypotheses, have achieved success in identifying clinically relevant associations. For example, EHRs were used in conjunction with genetic data to validate associations between single nucleotide polymorphisms and specific diseases.

As the authors of SeaWAS admitted, EHR observational studies have limitations—the existence of bias in health care data is well-known and well-documented. Comparison of large-scale EHR research with gold-standard, manually curated research has demonstrated that the two can produce inconsistent results. Known examples of systemic biases in EHRs include selection bias, coding bias, and missing or inaccurate records. In addition to EHR biases, retrospective studies in general may overstate effects and be subject to confounding factors. However, EHR research has been shown to provide an approximation to traditional research and has had several successful replications of large studies.

Boland, et al.’s hypothesis-free methodology was able to confirm known disease-birth month associations, discover new associations, and find clusters of birth month dependencies among disease types. Their work was significant because it applied sophisticated statistical techniques to a large dataset to derive novel insights. However, their work was limited by the cohort to which they applied their method. The SeaWAS study investigated records of 1.7 million individuals at New York-Presbyterian/Columbia University Medical Center. Observed associations from this population may be local effects that do not generalize to a more general population.

We conducted a retrospective study to apply the SeaWAS approach to larger cohorts from different geographical regions. By simultaneously applying this methodology to three separate cohorts, we were able to discern the effect of geographical, administrative, and population-based cohort differences. The increased sample size of our cohorts...
allowed increased power in detecting birth month associations. In addition, we proposed a change to the methodology used by SeaWAS to address statistical concerns first raised by Boland, et al and tested the impacts of its implementation.

Methods
Data Preparation
The individuals of interest were derived from the Explorys platform\(^3\). Patients were separated into three cohorts based on ZIP3 codes corresponding to regions in three different states. The first cohort (C1) consisted of patients in a southern US state at approximately 31° N latitude. The second cohort (C2) was constructed of patients from a midwestern US state around 40° N. The third cohort (C3) included patients from a western US state at approximately 38° N. Each of these cohorts represents an aggregation of multiple clinical and claims data sources, grouped by patient. In order to match patients across data sources, demographic records were matched on date of birth, gender, ZIP3, and the New York State Identification and Intelligence System (NYSIIS)\(^2\) representation of the patient’s name. All records were de-identified prior to analysis and all records derived from Centers for Medicare & Medicaid Services (CMS) data were excluded from the study.

Demographic information about both the original and replication cohorts is summarized in Table 1.

Table 1. Demographic information about original and replication cohorts.

<table>
<thead>
<tr>
<th></th>
<th>SeawAS (NY)</th>
<th>Replication (C1)</th>
<th>Replication (C2)</th>
<th>Replication (C3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Individuals [ Count ]</strong></td>
<td>1,749,400</td>
<td>4,588,300</td>
<td>4,840,500</td>
<td>2,331,000</td>
</tr>
<tr>
<td><strong>Sex [ Count (%) ]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>956,465 (54.67)</td>
<td>2,598,592 (56.64)</td>
<td>2,597,027 (53.65)</td>
<td>1,275,957 (54.73)</td>
</tr>
<tr>
<td>Male</td>
<td>791,534 (45.25)</td>
<td>1,988,451 (43.34)</td>
<td>2,240,720 (46.29)</td>
<td>1,054,662 (45.25)</td>
</tr>
<tr>
<td>Other/unidentified</td>
<td>1,401 (0.08)</td>
<td>1,289 (0.02)</td>
<td>2,757 (0.06)</td>
<td>362 (0.02)</td>
</tr>
<tr>
<td><strong>Race [ Count (%) ]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>665,366 (38.03)</td>
<td>2943136 (64.14)</td>
<td>3,087,731 (63.79)</td>
<td>1,563,300 (67.07)</td>
</tr>
<tr>
<td>Other/unidentified</td>
<td>842718 (48.18)</td>
<td>951485 (20.74)</td>
<td>983,021 (20.31)</td>
<td>623,719 (26.76)</td>
</tr>
<tr>
<td>Black</td>
<td>189,123 (10.81)</td>
<td>553,657 (12.07)</td>
<td>752,837 (15.55)</td>
<td>56,987 (2.44)</td>
</tr>
<tr>
<td>Declined</td>
<td>29,747 (1.70)</td>
<td>350,33 (0.76)</td>
<td>48,903 (1.01)</td>
<td>38,336 (1.64)</td>
</tr>
<tr>
<td>Asian</td>
<td>20,746 (1.19)</td>
<td>69,909 (1.52)</td>
<td>48,523 (1.02)</td>
<td>38,162 (1.64)</td>
</tr>
<tr>
<td>Native American/Indian</td>
<td>1,511 (0.09)</td>
<td>15,704 (0.34)</td>
<td>9,147 (0.19)</td>
<td>10,486 (0.45)</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>189 (0.01)</td>
<td>19,376 (0.42)</td>
<td>971 (0.02)</td>
<td>0 (0.00)</td>
</tr>
<tr>
<td><strong>Ethnicity [ Count (%) ]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>590,386 (33.75)</td>
<td>2,831,416 (61.71)</td>
<td>2,559,409 (52.87)</td>
<td>1,282,259 (55.01)</td>
</tr>
<tr>
<td>Unidentified</td>
<td>458,071 (26.18)</td>
<td>1,177,856 (25.67)</td>
<td>2,066,994 (42.70)</td>
<td>693,478 (29.75)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>361,123 (20.64)</td>
<td>560,298 (12.21)</td>
<td>83,991 (1.74)</td>
<td>277,226 (11.89)</td>
</tr>
<tr>
<td>Declined</td>
<td>339,820 (19.42)</td>
<td>18,730 (0.41)</td>
<td>1,301,06 (2.69)</td>
<td>78,037 (3.35)</td>
</tr>
<tr>
<td><strong>Other [ Median (IQR†) ]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>38 (22-58)</td>
<td>45 (27-62)</td>
<td>49 (29-66)</td>
<td>46 (29-64)</td>
</tr>
<tr>
<td>Years of follow-up</td>
<td>1 (1-3)</td>
<td>1 (0-3)</td>
<td>3 (0-7)</td>
<td>1 (0-3)</td>
</tr>
</tbody>
</table>

† Interquartile range

A list of diagnoses was derived from each patient’s medical history documents, problem lists, billing records, and other clinical findings. All patient records that contained an International Classification of Diseases, version 9 or version 10 (ICD-9, ICD-10) code were aggregated. Using a custom map based on the Common Data Model (CDM) mapping\(^8\) and Intelligent Medical Objects (IMO) data\(^4\), ICD concepts were converted to Systemized Nomenclature for Medicine-Clinical Terms (SNOMED) codes. The ICD to SNOMED map used in this study is not one-to-one; that is, an ICD code may map to more than one SNOMED code. In this case, all relevant SNOMED codes were recorded and included in analysis.

Statistical Methodology
For each SNOMED code with more than 1,000 distinct patients, a Pearson’s chi-squared test of independence\(^3\) was performed comparing the birth month distribution of patients with the condition to the birth month distribution of all
Boland, et al applied the Benjamini-Hochberg (BH) multiplicity correction to the p-values resulting from the chi-squared test. The BH multiplicity correction is a sequential hypothesis rejection procedure designed to control the False Discovery Rate (FDR) of independent test statistics. However, as Boland states, “Study limitations include the lack of condition independence […] potentially affecting multiplicity correction” (1051). We applied a more conservative multiplicity correction to our p-values, the sequential Holm multiplicity correction. To evaluate the impact of the change in multiplicity correction, we calculated the results using both multiplicity corrections and compared the output of the different methodologies.

**Results**

In all replication cohorts, we identified several literature-supported birth month-disease associations. We used the curated reference set of literature-supported associations from Boland, et al as a baseline to compare the results from the four cohorts. SeaWAS identified seven out of sixteen associations. Using the same multiplicity correction as the original study, the replication cohorts identified ten, eleven, and three associations at the adjusted p<.05 significance level. When considering associations only identified using the more conservative Holm correction, the replication cohorts identified nine, seven, and one of the literature-supported associations. The literature-supported associations identified in each cohort are given in Table 2.

<table>
<thead>
<tr>
<th>Literature-Supported Association</th>
<th>SeaWAS (NY)</th>
<th>Replication (C1)</th>
<th>Replication (C2)</th>
<th>Replication (C3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy/Asthma/Rhinitis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>(BH)</td>
</tr>
<tr>
<td>Reproductive Performance</td>
<td>X</td>
<td>X</td>
<td>(BH)</td>
<td></td>
</tr>
<tr>
<td>Eye Problems</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>(BH)</td>
<td>(BH)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory Syncytial Virus</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colitis</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leukemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADHD</td>
<td>X</td>
<td>X</td>
<td></td>
<td>(BH)</td>
</tr>
<tr>
<td>Atherothrombosis</td>
<td></td>
<td></td>
<td>(BH)</td>
<td></td>
</tr>
<tr>
<td>Atopic Dermatitis</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Fibrosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Otitis Media</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>X</td>
<td></td>
<td></td>
<td>(BH)</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 1 Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In addition to identifying correlations from the literature, we evaluated the previously unidentified associations discovered in each cohort. Out of the sixteen new associations from SeaWAS, four of them were replicated in at least one of the replication cohorts. All of these replicated findings were cardiovascular conditions. None of the non-cardiovascular associations discovered by SeaWAS were replicated. Out of the thirty-five unique conditions identified in an Explorys cohort, eight were replicated in another cohort. A summary of these findings, broken down by multiplicity correction and cohort, is presented in Table 3. The list of conditions identified is in Appendix A.
Discussion

In this replication attempt, we considered fifty data sources in three widely separated regions. The results differed greatly between different geographic regions. The most striking difference between cohorts in different geographical areas is the relatively few associations detected in cohort C3. Cohort C1 produced almost eight times more associations than did the C3. It has been shown that the strength of a birth month-disease association depends on the latitude, but birth month effects are typically stronger in higher latitudes. It is unclear why cohort C3, which is at a higher latitude than C1, produced fewer associations. It may be due to different health care processes used in different locations, or it may be related to differences in the characteristics of the individuals in the cohort. The variation in total number of associations observed and the fact that only six out of thirty newly observed associations were statistically significant in more than one cohort suggests that differences between cohorts may be a driver of substantial variation in the results of hypothesis-free searches.

Despite the lack of associations produced in C3, we have confidence that the replication cohorts were well-suited to test the hypothesis-free search. The C1 and C2 cohorts identified more of the literature-supported associations than did SeaWAS. This indicates that the sample size and extent of data was sufficient to detect legitimate birth month associations present in the cohorts.

In all of the cohorts tested, we failed to replicate the non-cardiovascular novel associations discovered by SeaWAS. These conditions include common ailments such as bruising, nonvenomous insect bite, vomiting, and venereal disease screening. Because these conditions are not typically very serious, their inclusion in an EHR may be dependent on the completeness of documentation and may suffer from seasonal bias.

Four new associations were discovered in the replication cohorts that were statistically significant in more than one geographical: coronary arteriosclerosis, tobacco use, hypoxemia, and reversible ischemic neurologic disease (RIND). The first association, coronary arteriosclerosis, may share a mechanism with the other cardiovascular conditions identified in SeaWAS. Although tobacco use has no obvious biological connection to birth month, the connection may be cultural, related to the impact of relative social age (as in Halldner or Skirbekk). The association of hypoxemia and birth month is mostly seen in individuals younger than five, suggesting that the observed effect may be transient. We are not aware of any explanation for the observed association of RIND with birth month.

The results discussed above strongly suggest that the Holm correction is an appropriate multiplicity correction to use for this hypothesis-free search. Although the recall of the search was higher with the Benjamini-Hochberg correction, there were many more non-replicable potential false positives identified using that correction. In addition to the non-replications of SeaWAS associations, the BH correction produced non-replicated associations in the new cohorts such as constipation, diaper rash, tinnitus, and headache. There is naturally a trade-off in choosing any threshold for significance; we suggest that the Holm multiplicity correction more accurately represents the uncertainties involved in this context.

This study had several limitations that could be overcome with future work. All the cohorts considered were in the northern hemisphere and the United States. Results from different regions of the globe may reveal different cultural

<table>
<thead>
<tr>
<th>Number of Literature Supported Associations Identified [ # (% Recall) ]</th>
<th>Number of Novel Associations Validated in Other Cohort [ # (% Precision) ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NY</td>
<td>C1</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Holm</td>
<td>-</td>
</tr>
<tr>
<td>BH</td>
<td>7 (44)</td>
</tr>
</tbody>
</table>

Table 3. Comparison of results for Holm and Benjamini-Hochberg (BH) multiplicity corrections. SeaWAS results are given in columns labeled “NY”; replication results given in “C1,” “C2,” and “C3” columns.
or environmental impacts on birth month associations. In addition, there were several confounding factors that we were unable to control in this retrospective study; two of which, age and coding differences between data sources, are known to impact studies. Future studies should more closely examine the impact of those confounding factors on the results of hypothesis-free searches.

**Conclusion**

This study is the largest of its kind with over 11 million unique patients, and is the first to include results from geographically distinct data sources. The size and breadth of this study allowed it to identify subtle trends and associations over large populations. We identified new birth month-disease associations and showed that results from a previous hypothesis-free search may not be generalizable. We have shown that the Benjamini-Hochberg multiplicity correction may not be well suited to a hypothesis-free search with dependent test statistics and shown that the Holm multiplicity correction is a better choice. It is clear that this approach can be a powerful hypothesis generator and tool for investigating the role of seasonally dependent early developmental mechanisms in general health, but obtaining generalizable results requires evaluating different sets of data and accounting for potential biases.

**Acknowledgements**

The feedback and assistance of the Innovations Team of Explorys, an IBM company was invaluable in preparing this work. I especially wish to thank Matthew Pohlman for his support throughout the research process and Amanda Yoho and Yifan Xu for their comments on the manuscript.

**References**

35. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag. 1900;5(50):157-175.
### Appendix A: List of conditions with statistically significant associations with birth month

<table>
<thead>
<tr>
<th>SeaWAS Description</th>
<th>p-value (BH)</th>
<th>Replicated?</th>
<th>SNOMED ID</th>
<th>Replication Description</th>
<th>p-value (Holm)</th>
<th>Validated?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial fibrillation</td>
<td>&lt;.001</td>
<td>Yes</td>
<td>55822004</td>
<td>Hyperlipidemia</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Essential hypertension</td>
<td>&lt;.001</td>
<td>Yes</td>
<td>387712008</td>
<td>Jaundice</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Congestive cardiac failure</td>
<td>&lt;.001</td>
<td>Yes</td>
<td>53741008</td>
<td>Coronary arteriosclerosis</td>
<td>&lt;.001</td>
<td>Yes</td>
</tr>
<tr>
<td>Angina</td>
<td>&lt;.001</td>
<td>No</td>
<td>399269003</td>
<td>Arthropathy</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Cardiac complications of care</td>
<td>0.027</td>
<td>No</td>
<td>90708001</td>
<td>Kidney disease</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Cardiomyopathy</td>
<td>0.009</td>
<td>No</td>
<td>92065004</td>
<td>Neoplasm of colon</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Pre-infarction syndrome</td>
<td>0.036</td>
<td>No</td>
<td>89765005</td>
<td>Tobacco use</td>
<td>&lt;.001</td>
<td>Yes</td>
</tr>
<tr>
<td>Chronic myocardial ischemia</td>
<td>0.022</td>
<td>No</td>
<td>40930008</td>
<td>Hypothyroidism</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Mitral valve disorder</td>
<td>0.024</td>
<td>Yes</td>
<td>7343006</td>
<td>Sleep apnea</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Acute upper respiratory infection</td>
<td>&lt;.001</td>
<td>No</td>
<td>43339004</td>
<td>Hypokalemia</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Bruising</td>
<td>0.015</td>
<td>No</td>
<td>93796005</td>
<td>Malignant neoplasm of female breast</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Nonvenomous insect bite</td>
<td>0.001</td>
<td>No</td>
<td>198036002</td>
<td>Impotence</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Venereal disease screening</td>
<td>0.003</td>
<td>No</td>
<td>201101007</td>
<td>Keratosis</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Primary malignant neoplasm of prostate</td>
<td>0.002</td>
<td>No</td>
<td>22325002</td>
<td>Abnormal gait</td>
<td>&lt;.001</td>
<td>No</td>
</tr>
<tr>
<td>Malignant neoplasm of overlapping lesion of bronchus and lung</td>
<td>0.014</td>
<td>No</td>
<td>414916001</td>
<td>Obesity</td>
<td>0.003</td>
<td>No</td>
</tr>
<tr>
<td>Vomiting</td>
<td>0.029</td>
<td>No</td>
<td>193462001</td>
<td>Insomnia</td>
<td>0.001</td>
<td>No</td>
</tr>
<tr>
<td>Radiculitis</td>
<td>0.001</td>
<td>No</td>
<td>2169001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharyngitis</td>
<td>0.001</td>
<td>No</td>
<td>363746003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypoxemia</td>
<td>0.002</td>
<td>Yes</td>
<td>389087006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.002</td>
<td>No</td>
<td>48694002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diarrhoea</td>
<td>0.007</td>
<td>No</td>
<td>62315008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polyneuropathy</td>
<td>0.013</td>
<td>No</td>
<td>42345000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delirium</td>
<td>0.016</td>
<td>No</td>
<td>2776000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiomegaly</td>
<td>0.024</td>
<td>No</td>
<td>8186001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acne</td>
<td>0.027</td>
<td>No</td>
<td>11381005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carpal tunnel syndrome</td>
<td>0.028</td>
<td>No</td>
<td>57406009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIND syndrome</td>
<td>0.028</td>
<td>Yes</td>
<td>36179005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sciatica</td>
<td>0.032</td>
<td>No</td>
<td>23056005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypoglycemia</td>
<td>&lt;.001</td>
<td>No</td>
<td>52767006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Improved Veteran Access to Care through the Veteran Health Information Exchange (VHIE) Retail Immunization Coordination Project

Nathan Botts, PhD\textsuperscript{1,2}, Eric Pan, MD\textsuperscript{1,2}, Lois Olinger, MA\textsuperscript{1,2}, Margaret Donahue, MD\textsuperscript{2}, Nelson Hsing, ScD\textsuperscript{2}

\textsuperscript{1}Westat, Rockville, MD; \textsuperscript{2}US Department of Veterans Affairs, Washington, DC

Abstract

The U.S. Department of Veterans Affairs (VA) Veteran Health Information Exchange (VHIE, formerly Virtual Lifetime Electronic Record, or VLER) Retail Immunization Coordination Project established a partnership between VA and Walgreens to empower Veterans to elect to receive their immunizations at a local Walgreens, which might be located closer to their home than their nearest VA facility. Analysis of Veterans immunized at Walgreens between September 2014 and January 2015 showed that 64\% of study Veterans now traveled <5 miles to receive their immunization, 12\% of study Veterans traveled between 5 to 10 miles, and 24\% of study Veterans traveled more than 10 miles. In addition, we note that 93\% of Veterans traveled less than 54 miles, the average distance rural Veterans traveled to the nearest VA facility. We conclude that the VHIE Retail Immunization Coordination Project improved Veteran access to healthcare and discuss future directions of this effort.

Introduction

VA influenza immunization rates for enrolled Veterans aged 65 years or older are currently above national averages (e.g., in 2011/2012, VA immunization rate was 77\% as compared to 65\% for the general population according to VHA Public Health data). However, VA remains below its target objective of 90\%. Many Veterans are currently already receiving immunizations through non-VA providers (e.g., employers, health outreach fairs, other clinics) and self-report their immunization history throughout the year to VA. With immunizations available at retail providers such as Rite Aid, Walgreens, CVS, etc. growing in popularity, VA is interested in determining how the convenience of retail providers, combined with automatically entered health information through interoperability, can be used to continue to improve the completeness and accuracy of the Veteran’s electronic health record and VA’s ability to accurately measure immunization rates for the target population, while boosting immunization rates of the target population overall.

VA launched the Retail Immunization Care Coordination Program (hereafter the immunization program) in October of 2013. The immunization program was structured in three phases:

**Phase 1 (2013-2014):** View only, query (Data was manually synched up by onsite VA staff following a clinical reminder. Data was transmitted from Walgreens, but onsite staff had to populate the patient record with immunization data.) It is also structured as a pilot program limited to Veterans Integrated Service Network (VISN) 8 throughout the state of Florida.

**Phase 2 (2014+):** Immunization data are automatically populating the VA patient record in CPRS. Phase 2 also extending the immunization program to a national rollout.

**Phase 3 (TBD):** Data will be integrated with Immunization Applications.

The immunization program allowed Veteran patients the option of choosing a Walgreens retail pharmacy to get a flu shot. Walgreens is authorized to administer the influenza vaccinations to the patients who are VA health care-enrolled and who voluntarily choose to participate in the program. In addition, many Veterans have coverage through Medicare or commercial insurance.

Immunization information of Veterans is available from Walgreens locations and retrieved for VA providers to review through VLER Health and eHealth Exchange, using the existing *query and retrieve* capability of VLER Health Exchange, and incorporates standard security measures to protect patient confidentiality. Participating Veterans in the VA Retail Immunization Care Coordination Program can ensure that their VA providers have complete and accurate information about the flu shot they receive, including the date, the type of vaccination, and the lot number. Once the Veteran volunteers to participate, the vaccination record is transmitted to the patient’s VA

326
electronic health record and is available for viewing via the VLER Health Exchange. VA providers will no longer have to rely on patient self-reported immunization data or unreliable paper records.

The goals of these efforts are to improve the completeness and accuracy of the Veteran’s electronic health record and to improve VA’s ability to accurately measure immunization rates for the target population, while boosting immunization rates of the target population overall. The immunization program expects to see both administrative impacts and clinical impacts from coordinating influenza vaccination with retail pharmacies (initially Walgreens).

Administratively, the retail influenza vaccination data will improve VA clinical documentation. This will eliminate the clinical reminders shown to VA clinicians for Veterans who were already vaccinated at retail sites such as Walgreens, and reduce cognitive load and alert fatigue to clinicians. Also, this will reduce unnecessary mails and outreach to Veterans for immunization reminders and self-reports once they are immunized. There will also be administrative savings from avoiding these wasted outreach activities.

Clinically, the program will provide real time influenza vaccination data to VA, which will help focus public health efforts to target Veterans who have not been immunized, reducing reminder fatigue and waste for Veterans who were already immunized. In addition, this is expected to add a large number of influenza vaccination sites and increased convenience for Veterans.

Finally, with greater and earlier influenza vaccination, there will be healthcare utilization impacts with reduced influenza-related phone calls, outpatient visits, and admissions. Influenza laboratory testing will also be reduced.

Objective

This paper covers the initial analyses of phase 2 of the immunization project, as the scope expanded nationally. Goals for the immunization pilot include the following: increased access to vaccinations by Veterans, enhanced vaccination rates, and improved data collection. Internal analysis of the Phase 1 results suggested that the immunization project is improving Veteran access and reaching a new population of Veterans. Therefore, Phase 2 analyses focused on confirming, identifying, and quantifying the impact on Veteran access. Phase 2 analyses leveraged Geographic Information System (GIS) to analyze geographic distribution of Walgreens sites providing immunization services, existing VA facilities, and residence of affected Veteran populations. The goal is to assist VA leaders and stakeholders to visualize how the immunization project is expanding Veteran Access.

Methods

We considered all Walgreens claims submitted to VA for the months of September 2014 through March 2015. The claims were matched against VA’s master patient index to confirm Veteran eligibility using patient matching algorithms and criteria according to established VA practices. Only claims with a confirmed “match” were considered in this analysis.

For each confirmed claim, travel distance was calculated between the Veteran’s home and the Walgreens store administering the immunization. The street addresses of the Walgreens stores were provided by the Walgreens national office. Due to Veteran privacy concerns, only the Veteran’s home zip code reported on the Walgreens claim was available for analysis. Therefore, travel distances were calculated from the centroid of the Veteran’s home zip code.¹ The ArcGIS Desktop program² from the Environmental Systems Research Institute (ESRI) was used to calculate the straight-line distance from the centroid of the Veteran’s home zip code to the Walgreens location where they received their immunization. Straight-line, or Euclidean distance is the minimum distance between points.

In addition, we also retrieved the location of the VA facility the Veteran frequents, based on the Veteran’s VA records and site ID, and calculated the corresponding travel distance using the same methodology as above. For Veterans who frequents more than one VA facility, the distances to each VA were calculated, and then averaged travel distance was used for the comparison.

To estimate the travel distances saved by Veterans, we only considered situations where the choice of being immunized at Walgreens led to a reduction in travel. Since these Veterans were eligible for free immunization at VA and had already received care at VA in the past (as shown by the presence of their medical records), any Veteran who traveled further than their usual VA facility to visit a Walgreens store further away were clearly not motivated to reduce travel distances, and had elected to visit Walgreens for a different reason. We also examined each Veteran’s VA visit records to determine whether they had VA visits during the influenza season before and after their Walgreens immunization.
Results

For the seven-month period between September 2014 and March 2015, there were 8,809 confirmed claims submitted by Walgreens for immunizing Veterans. Of these 8,809 claims, 827 (9.4%) had poorly formatted zip codes (neither 5 nor 9 digits), unused zip codes, or unused Walgreens store numbers. The remaining 7,982 claims were mapped to the Veteran’s home zip code (Figure 1) and used for the following travel distance analysis. Straight-line distances were calculated for each validated claim from the centroid of the Veteran’s zip code to the specific Walgreens location and the VA facilities the Veteran frequents. A further 473 (5.3%) claims where the Walgreens site visited was further away than the Veteran’s usual VA site of care. The results are shown in Figure 2. As can be seen from Figure 2, travel distances to Walgreens for these Veterans are usually 5 miles or less, whereas travel distances to VA facilities are usually 40 miles or more. After excluding Veterans who traveled to distant Walgreens for immunizations, who were presumably not motivated by reductions in travel distances, we found that during the 2014-2015 influenza season, these Veterans avoided a total of 2,055,240 miles by traveling to a near-by Walgreens store as opposed to their usual VA facility.

![Map of Veterans by 5-Digit Zip Code](image)

**Figure 1.** Number of Veterans by 5-Digit Zip Code, September 2014 to March 2015

When these Veterans’ VA visit records were review, 55% of the Veterans had one or more visits to VA facilities between the start of the influenza season on September 1, 2014 and the day when they were immunized at Walgreens. These were missed opportunities to vaccinate Veterans earlier. Further comparison of the Walgreen visits and VA visits were graphed below in Figure 3. Seventy-six percent (76%) of Veterans immunized at Walgreens visited a VA facility after their Walgreens immunization during the 2014-2015 influenza season. However, there was an average of 45 calendar days between their Walgreens immunization and the following VA visit; this represented days of extra protection against influenza and related illnesses even when compared to the ideal case that all Veterans were immunized at their next VA visit. Of the other 24% of Veterans who did not visit a VA facility during the influenza season after their Walgreens immunization, one percent (1%, the red slice in Figure 3) did visit VA prior to their Walgreens immunization; these were lost opportunities that VA never had a chance to correct. The remaining 23% never visited a VA facility during the entire 2014-2015 influenza season. It is quite possible that these 23% of Veterans would never have been immunized without this Retail Immunization Care
Coordination program. At the very least, it is certain that VA would never receive the immunization records of these Veterans without the Retail Immunization Care Coordination program, and they will erroneously remain on the unimmunized list for VA’s population health programs.

**Figure 2.** Veteran travel distances to VA versus Walgreens

**Figure 3.** Proportions of Veterans who visited VA before or after Walgreens
Discussion

This study applied GIS analysis techniques to the immunization claims submitted by Walgreens to VA in two ways: graphic visualization of geospatial distribution, and quantitative analysis of travel distances. Each claim used for this analysis represented a confirmed Veteran whose demographic information matched closely to the records in VA’s Master Veteran Index (MVI) database. In addition, the Veterans’ VA visit records were compared against their Walgreen immunization records to examine their chronological relationships. While the study is limited by the use of the Veteran’s home zip centroid and straight-line distance, these limitations applied equally to both Walgreens and VA facilities, so the impact on the travel distances compared should be minimal, and if anything, probably resulted in a conservative, lower-bound estimates of distance traveled.

Whereas Phase 1 of the immunization program was limited to VISN 8 in the State of Florida, the claims analyzed were from Phase 2 during national deployment. Visual inspection of Figure 1 confirmed the national nature of phase 2 of the immunization program, with activities shown in every state, including Alaska and Hawaii. In addition, it should be observed that several sites appeared to have higher activities than Florida, the original pilot sites. This confirmed the success of the Phase 1 pilot, and the resulting decision to deploy the immunization program nationally.

In addition, the distances traveled by these Veterans to receive immunization at Walgreens were estimated and detailed in Figure 2. It can be seen that the great majority of Veterans (73%) traveled more than 40 miles to their usual VA facilities, the distance under which the Veteran Access, Choice, and Accountability Act of 2014 (VACAA) used to define excess travel. While this is not a strict determination of VACAA eligibility, since VACAA used nearest VA facility whereas this study used the VA facilities the Veterans frequent, it nonetheless illustrated the significant travel burden for this group of Veterans. While this study did not include a direct comparison against the distances these Veterans would travel to the nearest VA medical facilities, comparison against published literature presented a very favorable picture. In 2006, Probst and her colleagues analyzed the 2001 National Household Travel Survey (NHTS) data collected by the US Department of Transportation, and found that national average distance traveled for medical and dental care was 10.2 miles. In 2010, Jeremy Mattson reported that the median distance traveled for healthcare services is 5 miles for routine health checkups, 9 miles for chronic healthcare visits, and 5 miles for emergency care. In contrast to the travel distance to VA facilities, only 7% of Veterans traveled more than 40 miles to Walgreens, seventeen percent (17%) traveled between 10 to 40 miles, fifteen percent (15%) traveled between 5 to 10 miles, and 60% traveled less than 5 miles. Using standard IRS mileage rates of 55 cents per mile for the estimated 2,055,240 miles saved by Veterans from travelling to a near-by Walgreens store for influenza vaccination, this is a savings of $1,182,000 ($0.55/mile x 2,055,240 miles) for Veterans. This study did not attempt to determine or estimate proportion of Veterans for whom VA provided transportation, travel assistance or reimbursements, which would further increase total travel cost and VA-specific costs.

As illustrated in Figure 3, 2,051 Veterans who never visited VA during the 2014-2015 influenza season were immunized at Walgreens, in addition, 6,908 Veterans were immunized earlier than their next VA appointment, gaining 323,328 days of protection against influenza and related illnesses. Based on the estimates reported by Duncan et al. in their cost benefits model, we estimated societal savings of $663.56 and $125.24 respectively for senior vs. adult Veterans who would not have been immunized otherwise. Using the same data set from the Duncan model, we also estimated the marginal benefit of earlier immunization at $3.16/day and $0.60/day respectively for senior vs. adult Veterans. Applying these estimated for the 2,051 Veterans (759 seniors, 1292 adults) who never visited VA during the 2014-2015 influenza season resulted in estimated societal savings of $665,452 (759 x $663.56 + 1292 x $125.24) and the 323,328 day of earlier influenza protection resulted in $499,492 (119,631 senior days x $3.16/day + 203,697 adult-days x $0.60/day) estimated savings.

Tallying the three sources of Veteran, VA, and societal savings from the Retail Immunization Care Coordination program during the 2014-2015 influenza season, we found a total estimated saving of $2,346,000. This compares quite favorable to the total VA expenditure for the 2014-2015 Retail Immunization Care Coordination program of $250,000, a mere 10.6% of the estimated savings.

Conclusion

By applying GIS analysis techniques, we were able to show how Phase 2 of VA’s Retail Immunization Coordination Program with Walgreens is improving Veteran access and furthering Veteran health. We first demonstrated how
Phase 2 of the immunization program had improved Veteran access by expanding the coverage beyond Florida to cover every state in the nation. We then estimated savings of $1,182,000 for Veterans from reducing more than 2 million miles in travel. We also estimated more than $1 million savings from earlier immunization and reaching unserved Veterans. We conclude that Phase 2 of the immunization program had improved Veteran access to healthcare, and VA should continue to expand the program with other private sector partners and additional services.

References


2. Mattson J. Transportation, distance, and health care utilization for older adults in rural and small urban areas. Fargo, ND: Small Urban & Rural Transit Center, Upper Great Plains Transportation Institute, North Dakota State University; 2010.

Automated Detection of Privacy Sensitive Conditions in C-CDAs: Security Labeling Services at the Department of Veterans Affairs

Omar Bouhaddou, PhD², Mike Davis, MS¹, Margaret Donahue, MD¹, Anthony Mallia³, Stephanie Griffin, JD¹, Jennifer Teal, RHIA¹, Jonathan Nebeker, MD¹

¹ U.S. Department of Veterans Affairs, Washington, DC; ²Hewlett Packard Enterprise, Plano, TX; ³Edmond Scientific, Barrington, NJ

Abstract
Care coordination across healthcare organizations depends upon health information exchange. Various policies and laws govern permissible exchange, particularly when the information includes privacy sensitive conditions. The Department of Veterans Affairs (VA) privacy policy has required either blanket consent or manual sensitivity review prior to exchanging any health information. The VA experience has been an expensive, administratively demanding burden on staff and Veterans alike, particularly for patients without privacy sensitive conditions. Until recently, automatic sensitivity determination has not been feasible. This paper proposes a policy-driven algorithmic approach (Security Labeling Service or SLS) to health information exchange that automatically detects the presence or absence of specific privacy sensitive conditions and then, to only require a Veteran signed consent for release when actually present. The SLS was applied successfully to a sample of real patient Consolidated-Clinical Document Architecture (C-CDA) documents. The SLS identified standard terminology codes by both parsing structured entries and analyzing textual information using Natural Language Processing (NLP).

Introduction
With increased care coordination across organizations and the need to comply with Meaningful Use requirements of the HITECH Act, more Health Information Exchanges (HIE) are happening, but providers are struggling to both share clinical information and remain in compliance with applicable privacy laws, as documented by the Office of the National Coordinator (ONC)¹. The Department of Veterans Affairs (VA) estimates that about 75% of Veterans receive part of their healthcare in the private sector and need information exchange. Title 38 U.S.C. § 7332 requires VA to obtain a signed authorization from Veterans whose health record contains one of four privacy sensitive conditions: diagnosis of HIV or sickle cell anemia, and diagnosis and treatment of drug and alcohol abuse, before disclosure to an outside entity². In the paper world, the VA Release of Information (ROI) offices manually review any request for health information, and assesses whether the information includes any privacy sensitive conditions (i.e., 7332-protected). If it does, then a signed authorization is obtained from the Veteran. If not, then the information is disclosed without patient authorization, as already permitted by the HIPAA Privacy Rule. In electronic HIE, no one is manually attending to the record sharing transaction which happens automatically over the eHealth Exchange network³ or when using the Direct Project⁴. VA’s current approach to comply with this requirement in the electronic exchange world is to always obtain a signed authorization, whether or not the information to be shared contains privacy sensitive conditions. This has been the strategy because the information disclosed is not reviewed manually and a reliable method for automatic detection of privacy sensitive conditions was not available.
While requiring that all Veterans sign an authorization is effective in meeting 38 U.S.C. § 7332 regulation, this universal opt-in model has placed an unnecessary burden on the majority of Veterans whose records do not contain these conditions and whose information could otherwise be exchanged without an authorization. Opt-In means that Veterans wishing to make their information available for exchange must actively express their privacy preferences by means of a signed written request (e.g., authorization). Over the last 5 years, the VA HIE program managed to collect 270,000 authorizations (3% of the 8.76 million actively enrolled Veterans). Today, 62% of the authorizations are collected on paper, while 38% are e-signed online. These authorizations are only valid for 5 years and 1/5 needs to be renewed each year. At this pace, to provide coverage for all enrolled Veterans, it would take 162 years! Finally, the low rate of Veterans participation contributes to the low rate of HIE transactions and missed opportunity for care coordination and outcome improvements.

This situation is not unique to VA. In addition to Title 38 U.S.C., there are many laws and regulations protecting the sharing of sensitive data, including HIPAA Privacy Rule, 42 CFR Part 2, Genetic Information Nondisclosure Act (GINA) of 2008, and many types of sensitive data, including Domestic Violence, Sexual Activity, HIV, STDs, Substance Abuse, Mental Health, Reproductive Health, VIPs, and Employees. Since reliable detection of privacy sensitive information has been considered beyond the capabilities of most HIEs, organizations tend to either adopt a universal opt-in consent model or simply do not share data from potentially sensitive categories.

This paper presents a proof of concept that challenges the current assumption. It assesses the reliability of a Security Labeling Service (SLS) that would automatically detect privacy sensitive conditions within the documents currently exchanged by VA with private sector healthcare providers. The format of these documents follows the Health Level Seven (HL7) standard Consolidated-Clinical Document Architecture (C-CDA) template, which is an XML format with a human-readable block (mostly text) and computable structured entries with coded information. The SLS complies with several HL7 standards including the Data Segmentation for Privacy. It parses coded information structured entries in C-CDAs and uses Natural Language Processing (NLP) for text blocks to distinguish records containing privacy sensitive conditions, requiring authorizations, from those that are not sensitive and do not require an authorization. The objective is to achieve effectively zero false negative (normal) results while at the same time minimizing false positive (restricted) results. The success of this proof of concept and follow-on studies could lead to substantial benefits by reducing the authorizations needed when no 7332-protected conditions exists. It would constitute a major milestone toward the ability to share millions of records automatically, increase provider adoption of HIE through less restrictive consent management for information that is not sensitive, and substantially reduce costly paperwork and burden on Veterans and VA staff.

Methods

The SLS uses a list of sensitive codes from standard terminologies (e.g., ICD-9, ICD-10, SNOMED CT) against which it compares the codes found in the C-CDAs. The codes in the C-CDAs are found either in the structured entries, as required by Meaningful Use Certification or are derived from NLP parsing of the textual information found in the narrative blocks and embedded reports in the C-CDAs. If any code found is part of the sensitive codes list, then the record is labeled ‘restricted’ and cannot be exchanged without patient authorization. If no data in the C-CDA is part of the sensitive codes list, then the record is labeled ‘normal’ and potentially could be exchanged without authorization.
Figure 1 describes the experimental design. A population study was identified from VA’s Corporate Data Warehouse (CDW) by selecting 90 patients with known privacy sensitive conditions and 30 at random (unknown). A C-CDA was generated for each patient and these C-CDAs were anonymized using a strict protocol. SLS scanned all 120 documents and produced a security label and explanations for each. A VA Release of Information (ROI) specialist reviewed the 30 ‘unknown’ documents and marked whether they were restricted or normal. The results were tabulated and analyzed by two of the authors.

Figure 1: Experimental design implemented in the SLS proof of concept.

The CDW query generated 40 for each protected condition but by the time we had generated the CCDAs (some failed to generate), we got a reduced control list of 32 HIV, 29 SCA and 32 Abuse which was 93 morbidities, but one record had two morbidities so there were 92 control records with morbidities. We also ended up with 32 unknowns.

SLS SOFTWARE DESCRIPTION
The SLS solution developed for this proof of concept is based on Health Level Seven (HL7) standards. Figure 2 below describes the SLS processing flow.

SLS parses and marks each C-CDA as normal or restricted. The SLS parsing is specific to the C-CDA Continuity of Care Document (CCD) template. The SLS scans the full document examining all structured entries for the presence of any privacy sensitive code. If one or more sensitive codes are found, then SLS documents what sensitive codes were found and what sections in the C-CDA they were found. Only if no sensitive codes are found does the SLS evoke the NLP engine to parse the text parts of the C-CDA (e.g., narrative blocks, embedded notes and reports). The NLP engine is the open source SPECIALIST NLP obtained freely from The National Library of Medicine. It parses the textual information in the C-CDA, identifies Unified Medical Language System (UMLS) codes by comparing words and word groups to UMLS concepts. The UMLS codes found are then expanded into their semantic mapping set and filtered down to the code systems that are part of the sensitivity codes list. These mapped codes are then compared to
Similarly to the code scan, the NLP scan does not stop at the first sensitive code found but parses all textual information in a C-CDA and identifies all privacy sensitive codes that exist.

**Figure 2**: SLS marks each C-CDA as restricted or normal and documents the sensitive codes found.

**Sensitivity Codes List**

This is a list of terminology codes representing the concepts of HIV, sickle cell anemia, drug and alcohol abuse in standard vocabularies, including ICD-9-CM, ICD-10-CM, HCPCS, CPT, SNOMED CT, LOINC, and RxNorm. A first draft of the sensitivity codes list was based on the publicly available SAMHSA list. VA Privacy and Health Information Management specialists refined this list by removing mental health conditions, since these are not ‘protected’ conditions within 38 U.S.C. § 7332, and by adding additional ICD-9-CM and ICD-10-CM codes missing. The list count is 3989 codes distributed across the standard coding nomenclatures as follows:

Table 1: Sensitive codes list distribution

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Code Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9-CM</td>
<td>510</td>
</tr>
<tr>
<td>CPT-4</td>
<td>56</td>
</tr>
<tr>
<td>HCPCS</td>
<td>6</td>
</tr>
<tr>
<td>ICD-10-CM</td>
<td>184</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>6</td>
</tr>
<tr>
<td>LOINC</td>
<td>1798</td>
</tr>
<tr>
<td>RxNorm</td>
<td>1429</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3989</strong></td>
</tr>
</tbody>
</table>
The sensitive codes list used in this proof of concept is relatively crude. Experience with the proof of concept has taught us the critical role it plays in sensitivity and specificity of the results. In the discussion section, we provide guidelines for a more robust creation, maintenance, and governance process of this list.

**Sample VA C-CDAs**

It was determined that a sample of 120 C-CDAs would be used for the proof of concept. The count of 120 represents more a practical convenience as our objective was not statistical significance. Another practical consideration led us to divide the cases in one group of known cases (90) and another group of unknown cases (30). A query was developed and run against the VA Corporate Data Warehouse to select these cases. The 90 known cases where identified as having one of the diagnoses specifically recorded as the reason for the visit in the last 12 months: 30 HIV, 30 sickle cell anemia (SCA), and 30 drug/alcohol Abuse. We did not make an attempt to confirm whether these 90 cases were true or whether the sensitive information was found in the codes, the narrative text or both. The 30 unknown cases were selected at random. The result was a table of 120 unique patient names and index numbers, where 90 entries had a ‘1’ under HIV, SCA, or Abuse column, and 30 entries, corresponding to the unknown cases, had their columns left blank, as illustrated in Table 2, CONTROL section.

A team of volunteers generated and anonymized a C-CDA for each patient. An anonymization task (not de-identification) was deemed sufficient to minimize privacy risks during processing within the VA EHTAC lab. The team manually anonymized the C-CDA documents using an XML editor. Specific rules were developed, tested, approved, and consistently implemented to remove names, date of birth, social security number, addresses, insurance IDs and contact information from structured and unstructured information contained in the C-CDAs.

Finally, the volunteers created 30 PDF files for ROI staff review, by rendering the 30 unknown C-CDAs using the VA C-CDA style sheet.

**Metrics measured**

For each case (i.e., C-CDA), results from SLS were tabulated against known conditions and ROI assessment, as illustrated in Table 2, SLS section.

| Table 2: Illustration of results table showing sensitive conditions identified in CONTROL and detected by SLS for the sample case ‘12345’. |
|-----------------|-------------|-------------|-----------|-------------|-------------|-------------|-------------|
| **CASE**       | **HIV**    | **SCA**    | **Abuse** | **Normal** | **HIV**    | **SCA**    | **Abuse**   | **Normal** |
| 12345          | 1          |            | 0         |            | 1          |            | 1           | 0          |

For the known cases, the objective is to compare SLS to control conditions. In this case, we can estimate sensitivity as the True Positive (TP) rate which is TP / all Positives. We are also interested to measure False Negative (FN) (i.e., False Normal) and any False Positives (FP). A FP creates the need for an unnecessary patient authorization but it is not as consequential as a False Negative (FN), which could result in a breach or unauthorized disclosure. The known cases were selected because they contained at least one privacy sensitive condition (restricted). However, we don’t know whether or not they contain additional privacy sensitive conditions, which is why the other columns in Table 2, CONTROL section were left blanks. Consequently, we cannot estimate specificity or True Negative (TN) rate.

For the unknown cases, the goal is to compare the SLS against ROI assessment. To calculate the overall percent agreement we take the total number of times in which SLS and ROI agree and divide that by the total number of classifications made. Because ROI staff assessment stopped at the first privacy sensitive condition found, we cannot
assess whether the SLS over detection are true positives, unless each case is reviewed by an impartial third-party and the detected conditions confirmed or denied.

Results
The proof of concept results are not expected to be statistically significant and no final decision is projected. However, we hope it would hold enough promise to pursue a more formal scientific study, upon which policy, business, and technical implementation decisions can be made.

Table 3 shows the exact count of Known conditions and detected conditions by SLS. Table 4 shows the exact count of conditions detected by ROI staff and SLS, in the Unknown cases. In both tables, SLS seems to have detected the sensitive conditions present, but also detected a few more. Table 5 and 6 take a closer look at these results, in terms of true positives and agreements.

Table 3: SLS conditions detection in Known cases.

<table>
<thead>
<tr>
<th></th>
<th>HIV</th>
<th>SCA</th>
<th>Abuse</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Cases</td>
<td>32</td>
<td>29</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>Detected by SLS</td>
<td>34</td>
<td>29</td>
<td>52</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4: SLS conditions detection in Unknown cases.

<table>
<thead>
<tr>
<th></th>
<th>HIV</th>
<th>SCA</th>
<th>Abuse</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROI</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>SLS</td>
<td>9</td>
<td>0</td>
<td>23</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 5: SLS sensitivity, ‘extra positive’ and FN rates within the Known cases. *Specificity cannot be calculated due to lack of knowledge of all privacy sensitive conditions in the Known cases.

<table>
<thead>
<tr>
<th></th>
<th>HIV</th>
<th>SCA</th>
<th>Abuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (TP rate)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Extra Positive</td>
<td>6%</td>
<td>0%</td>
<td>40%</td>
</tr>
<tr>
<td>Specificity (TN rate)*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN rate</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 5 shows that for every known case, SLS was able to detect the existence of the known condition in the case: HIV, SCA, or Abuse. Therefore, the SLS sensitivity for known positives is 100%. Also, for all known cases (all restricted), SLS never mislabeled a case as normal. Therefore, the FN rate is 0% and this is very good news as FN is the error we wanted to eliminate. However, SLS over detected many more conditions (HIV and Abuse) per case than were known as described in the section below. True FP and TN cannot be calculated, because we don’t know the status.
of all privacy sensitive conditions in the Known cases, unless we inspect them manually. Only TP and FN rates can be estimated. Consequently, specificity cannot be estimated.

**Table 6**: SLS agreement with ROI staff on Unknown cases.

<table>
<thead>
<tr>
<th></th>
<th>Restricted</th>
<th>Normal</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>8</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>Disagree</td>
<td>0</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Agree %</td>
<td>100%</td>
<td>36%</td>
<td>52%</td>
</tr>
<tr>
<td>Disagree %</td>
<td>0%</td>
<td>64%</td>
<td>48%</td>
</tr>
</tbody>
</table>

In Table 6, we note again that, every time ROI staff said a case was restricted, SLS agreed (100% agreement). This means that there was no case where ROI staff identified a privacy sensitive condition but SLS mislabeled it as normal. However, SLS detected additional sensitive conditions in several cases. Despite this over detection, out of about 75% Normal (25/33) determined by ROI, SLS found about 25% (9/33). This 25% is the low estimate of the auto opt-in that SLS could clear, in other words, of the benefits of SLS. Further, the over detection can be easily corrected, as discussed below.

**SLS OVER DETECTION OF FALSE POSITIVES**

There were several cases where SLS over detected an HIV condition that, after inspection, did not confirm as a true positive. This was due to the NLP limited interpretation of questionnaires and negations. Free-text questionnaires would often include discussion about patient being offered an HIV screening and declining. In other cases, Veterans accepted HIV testing and had negative results. SLS interpreted these lab results correctly and did not label them as positive. In one case only, an HIV misdiagnosis was made and later disputed by the patient and corrected in the notes. There was no FP in the case of sickle cell anemia. In the case of alcohol/drug abuse there were many FPs. Again, many were due to NLP of questionnaires and the handling of responses. Questions on alcohol use with answers of ‘no’, ‘never’, ‘occasionally’, etc. were counted as positive. Another contributor factor to FPs was the inclusion by mistake of ‘nicotine/tobacco use’ in the sensitive codes list.

**CODES VS. NLP**

The NLP was evoked only if there were no sensitive codes found in the C-CDA structured entries. As shown in Table 7 below, NLP was used only minimally. In other words, C-CDA structured entries provided the source of truth for the large majority of cases. SLS found multiple privacy sensitive conditions per record, particularly Drug/Alcohol abuse due to its systematic inclusion in questionnaires.

**Table 7**: Use of C-CDA structured entries codes vs. NLP of textual information.

<table>
<thead>
<tr>
<th></th>
<th>HIV</th>
<th>SCA</th>
<th>Abuse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLS detection overall</td>
<td>43</td>
<td>29</td>
<td>75</td>
<td>147</td>
</tr>
<tr>
<td>SLS detection by code only</td>
<td>34</td>
<td>27</td>
<td>65</td>
<td>126</td>
</tr>
<tr>
<td>SLS detection by NLP (when code finds none)</td>
<td>9</td>
<td>2</td>
<td>10</td>
<td>21</td>
</tr>
</tbody>
</table>
Because our relatively simple approach to NLP, we had many false positives. We examined what the SLS results would be without the NLP. We found that SLS would ‘miss’ 5 sensitive conditions: 3 HIV and 2 sickle cell anemia. Are these conditions really false negatives? Upon closer examination, we found explanations for each case, as described below:

- **HIV cases**: the first case was selected because of the presence of the ICD-9 code for Kaposi’s sarcoma. This code was not included in the sensitive codes list, which explains why the code search did not find it. The NLP found a lab test of ‘HIV SCREEN Ag/Ab’ with a result of ‘Nonreactive’, which it counted as positive. NLP needs to do better with textual lab test results. The second case was similar to the first, except that NLP labelled it positive because of a questionnaire item of ‘Have you tested positive to HIV? No’. The third control case had the ICD-9 code 079.51 HUMAN T-CELL LYMPHOTROPHIC VIRUS. This code was missing from the sensitive codes list, which is why a code scan could not find it. NLP found it because of the question ‘HIV Status: Unknown’.

- **Sickle cell anemia cases**: The first case was selected as a control for a reason (code) yet to be determined. The C-CDA did not include a coded entry in the problem section for sickle cell anemia. The NLP labelled it as sickle cell anemia because it found a ‘provider comment’ of ‘sickle cell trait’ associated to the non-specific code ‘Anemia (ICD-9-CM 285.9)’. In the second case, an independent review of the patient complete medical record confirmed the presence of sickle cell anemia (ICD-9 282.5) diagnosis. However, the C-CDA did not include this entry in the problem section for a reason yet to be determined. This explains why SLS code scan did not find it. However, NLP found it because of this entry in the problem list: ‘sickle cell trait (ICD-9-CM 799.9)’. ICD-9 code 799.9 is not a sensitive code because it is a generic code used to allow users to enter a free-text entry when no appropriate ICD-9 code is found in the problem list.

**Discussion**

We noted considerable improvements with successive refinements of the sensitive codes list, over the multiple runs of the SLS. Going forward, it is imperative that a more concerted effort be given to the creation, maintenance and governance of this list. This effort must bring together a governance board including clinicians, privacy, medical record coding, terminologists, security architects, interoperability specialists, and others. It would take into account that certain diagnoses are markers for others (e.g., Kaposi’s sarcoma and HIV). Some lab tests can be ordered for multiple reasons. Medications can have multiple indications and out of label use (e.g., acyclovir). Diagnoses are more specific than lab tests or medications. Finally, it would be created from existing lists, and eventually converge toward a well-defined standard terminology subset.

Also, we found one particular quality issue with the EHR data. The code ‘042.’ was associated with a diagnosis of HIV and missed by SLS in the early runs, as this code is not a valid ICD-9-CM code and was not in the sensitive codes list. The valid code is ‘042’ without the period. To address this issue quickly, we added the code ‘042.’ to the sensitive codes list. Also, we were careful with the use of spreadsheets to manage terminology codes, in particular in the case of ICD-9-CM codes because leading and trailing zeros can disappear in spreadsheets if cells are not formatted appropriately to allow for that (i.e., format as text rather than numbers).

Although the role of NLP was minimal given the extent to which VA C-CDA’s have coded data, the challenges of questionnaires and negations need to be better managed if we want to minimize false positives. We understand from experts in the NLP field that more sophisticated NLP engines address these issues and other studies have shown the benefits of combining NLP with scanning for codes to collect supplemental information. We plan to use these refined tools in the future.
The sample set was too small and not representative of the Veteran patient population to draw any conclusion from it. In the next phase of the project, we need a sample where all conditions are identified in a case so that both sensitivity and specificity of the SLS can be estimated.

Conclusion
Reliable automated detection of 38 U.S.C. § 7332 protected conditions in VA C-CDAs is possible! The SLS proof of concept was successful. Sensitivity of the detection of known cases was optimal and there were no false normal, meaning, there would be no risk of a breach because of a disclosure without authorization when required. This initial study may change the way VA handles electronic HIE disclosures in the future. However, before such a business and policy decision can be made, a larger and more rigorous study is needed. We are in the process of designing such a study that will be based on a larger sample of C-CDAs and would take place in the VA production environment, alleviating the need for anonymization. There are sophisticated commercial security labeling products, using advanced NLP and comprehensive sensitivity code lists. The objectives would be twofold: 1) confirm the reliability of the SLS performance, meaning high sensitivity and low or no false normal, and 2) estimate the percent of cases that can be reliably cleared through this automatic detection process and the benefits associated. If successful, then the VA can integrate SLS and electronic HIE. An SLS could also become a support tool for the manual ROI process. Finally, beyond Title 38 U.S.C., there are other labeling that can benefit from SLS such as VA Employees, VIP records and Meaningful Use 2016 security labeling.

Acknowledgements
This work could not have been possible without the contribution of many, including Charles Demosthenes, Augie Turano – VA Corporate Data Warehouse; Jamie Bennett, LeAnn Roling, Jen Cockle, Monique Allen, Phyllis Denson, Melissa Sands, Marie Swall– VHA Interoperability Office; Mohammad Jafari, Kathleen Connor, Duane Decouteau – VHA Security Architecture; Margaret Pugh – VHA Privacy; Barbara Mulvaney and Barbara Freeby – Richmond VA Medical Center.

References
4. Direct Project (initially developed by the Office of the National Coordinator) - http://directproject.org/
8. Substance Abuse and Mental Health Services Administration. CENTER FOR INTEGRATED HEALTH SOLUTIONS http://www.integration.samhsa.gov/operations-administration/HIE_paper_FINAL.pdf
Safe “cloudification” of large images through picker APIs

Erich Bremer¹, Tahsin Kurc¹, Yi Gao¹, Joel Saltz¹, Jonas S Almeida¹*

1) Dept Biomedical Informatics, Stony Brook University (SUNY), NY 11794
* jonas.almeida@stonybrookmedicine.edu

Abstract

The “Box model” allows users with no particular training in informatics, or access to specialized infrastructure, operate generic cloud computing resources through a temporary URI dereferencing mechanism known as “drop-file-picker API” (“picker API” for sort). This application programming interface (API) was popularized in the web app development community by DropBox, and is now a consumer-facing feature of all major cloud computing platforms such as Box.com, Google Drive and Amazon S3. This reports describes a prototype web service application that uses picker APIs to expose a new, “cloudified”, API tailored for image analysis, without compromising the private governance of the data exposed. In order to better understand this cross-platform cloud computing landscape, we first measured the time for both transfer and traversing of large image files generated by whole slide imaging (WSI) in Digital Pathology. The verification that there is extensive interconnectivity between cloud resources let to the development of a prototype software application that exposes an image-traversing REST API to image files stored in any of the consumer-facing “boxes”. In summary, an image file can be upload/synchronized into a any cloud resource with a file picker API and the prototype service described here will expose an HTTP REST API that remains within the safety of the user’s own governance. The open source prototype is publicly available at sbu-bmi.github.io/imagebox.

Availability

The accompanying prototype application is made publicly available, fully functional, with open source, at http://sbu-bmi.github.io/imagebox. An illustrative webcasted use of this Web App is included with the project codebase at https://github.com/SBU-BMI/imagebox.

Introduction

The advantages of user-facing cloud computing are particularly hard to realise for specialized image analysis applications because of the dependence on specialized libraries and infrastructure, as is the case for the popular Open Slide platform [Goode 2013]. The challenge of presenting image analysis solutions to those that do not have those specialized resources is significantly compounded if they target large images in private stores. In other words, image informatics applications are currently dependent on specialized informatics infrastructure operated by specialized human resources. This is of course not to the demerit of the specialized
libraries themselves, such as the excellent OpenSlide, it instead reflects a missing component in the cloud computing landscape. Two of the most compelling scenarios where these challenges arise are a) when one or more regions in a whole slide image needs to be object of a specialized analysis; and b) when a private image needs to presented to a domain user that does not have access to web-facing image analysis infrastructure (which is the norm rather than the exception).

The immediate motivation for this work is to address requirements associated with the set of intense research efforts that focus on linking morphologic and molecular cancer characterizations and to make use of combined imaging and molecular signatures to substratify patient populations to predict outcome and response to treatment. A number of reports by the National Cancer Institute describe efforts being carried out by the pathology and radiology research communities to link imaging phenotypes with large-scale genomic analyses, ultimately informing personalized medicine clinical trial designs such as NCI-Match (https://clinicaltrials.gov/ct2/show/NCT02465060). Characterization of tissue morphology related phenotypes is a complex effort that requires the development and deployment of a variety of pipelines, many of which require flexible access to image sections. Accordingly, the cloud API mediator described here was developed in the context of Pathology images. However, this “cloudification” approach is equally applicable to other image analysis infrastructure scenarios such as those involving Radiology.

These challenges of commoditized and portable biomedical informatics analysis are of course not a problem exclusive of image analysis applications. Novel REST API architectures are indeed emerging to address them as a generic solution. Specifically, cloud computing infrastructure has responded with new interoperability models [ONC 2015], as illustrated by the central role of the HL7 FHIR API (https://www.hl7.org/fhir), which create novel opportunities for architecturing distributed whole slide image processing. These opportunities for RESTful interoperation are assessed in this report through the use of the accompanying webApp application prototype, which places no demands on the user beyond the client Web Browser. In other words, the functionality of the “cloudified” image traversal API it validated by a prototype Web Application. This application will seek to engage the “Box / file picker” application programming interface (API) exposed by a Box cloud resource, which have emerged to interoperate with infrastructure of a wide variety of cloud providers. Picker APIs, after successful authentication (typically using OAuth 2.0), expose temporarily dereferenceable URLs which can be picked up by cloud-based applications in order to make full use of the high connectivity that characterizes hosting data centers.

The development of high performance “slice servers” (which extract slices of different regions, at different scales, from large image files) has also advanced significantly, as illustrated by the open source IIP project (iipimage.sourceforge.net). Furthermore, the ability to approach image analysis either as a web service or as computation entirely on the client side has also been object of considerable attention. The former is typically a commercial application, such as Aperio’s ePathology product [Leica 2015], and the latter is typically an Academic exploit, such
as our own [Almeida 2012]. As a result of these advances in portable image presentation, the obstacles to a consumer-facing approach to interact with large images, wherever they may be, is now reduced to the enablement of an easily distributable programmatic interoperability mechanism.

These developments place the picker API at an interesting intersection between cloud computing and high performance slice servers. As regards Biomedical applications, this intersection has, however, been associated with lack of sufficient security and privacy to allow users such as Pathologists and translational researchers. This final obstacle to exploring cloud (HTTP REST) API implementation architecture was also removed in recent times by the adoption of HIPAA compliant practices by a number of cloud providers, such as Box.com [Box 2015], Google Drive [Google 2015], and Microsoft OneDrive [Microsoft 2015].

Methods

ImageBox WebApp

The imageBox Web App was developed entirely as a within browser application, using only the Web’s “assembly language”, JavaScript. This approach, which relies exclusively on the Web Platform (webplatform.org) and W3C open standards, leads to applications that are assembled directly in the web browser of the user, with no requirement for software download and installation, concern about native library dependencies or exposure of sensitive data to third parties. As a consequence, the imageBox Web App was developed in the public domain with open source and version control from its onset (https://github.com/SBU-BMI/imagebox). The fully functional application itself is served directly from GitHub by maintaining the application code in the “gh-pages” branch (i.e. live Web App served at sbu-bmi.github.io/imagebox). This signifies that the imageBox Web App is being delivered with versioned hosting, such that sites like https://rawgit.com can expose individual versions programatically.

TCGA images

The operation of ImageBox was tested for whole slide images of The Cancer Genome Atlas. These images were both called directly from TCGA’s public web directory, or where downloaded to private Box.com and DropBox folders. The latter served the purpose of validating the desired alignment of imageBox with the governance of private whole slide image files without compromising the content of the hosting folder (see File-picker API below and also Discussion). At the time of writing this report, the public folder with tens of thousands of TCGA images for all 30+ tumor types can be found at https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/<tumor-type>/bcr/nationwidechildrens.org/tissue_images/slide_images/, using the codes for <tumor-type> listed at https://tcga-data.nci.nih.gov/tcga. For example, for Glioblastoma Multiforme, the TCGA slide image folder would be at
https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/bcr/nationwidechildrens.org/tissue_images/slide_images. However, NCI has announced the intention of moving access to public TCGA data to the new Data Commons resource at https://gdc-portal.nci.nih.gov, so in the future the URL composition may need to be redirected to that new resource.

File-picker API

The implementation of the file-picker API mechanism varies between “Box providers” (cloud resources that expose file-picker APIs, see Introduction) even if they deliver the same final result: a uniquely dereferenceable URL. The two Boxes tested by imageBox are Box.com and DropBox. The details of their file-picker APIs can be found, respectively, at https://developers.box.com/the-box-file-picker and https://www.dropbox.com/developers/dropins/chooser/js.

Cloud-hosted imageBox slicer

This component, represented as a cloud box and labeled “imageBox” in Figure 1, uses the Jetty server (http://www.eclipse.org/jetty) to handle whole-slide microscopy images upwards of 100,000 pixels by 100,000 pixels. To enable the imageBox tiling server to access the SVS image files, the Bioformats toolbox (http://www.openmicroscopy.org/site/products/bio-formats) was used, which was developed to read and write image data between different formats with a focus on digital pathology. As detailed in the project’s documentation page, this toolbox relies on the OME-Model (http://www.openmicroscopy.org/site/support/ome-model) to slice and scale image pyramids, and is then exposed by ImageBox API. To retrieve the OME annotation of a imageBox link, add “&format=json” and the end of a slice URL, which will otherwise return a pixel map.

Results

A prototype application (see Availability) was developed to engage a image tiling server (“slice server”, see Methods) as part of a file-picking API call to private cloud storage. This application was used to measure transfer and retrieval rates for both moving the full SVS file to a well connected (cloud) slice server, and then retrieving image sub-regions at different scales. As schematized in Figure 1, this cloud-based “imageBox” resource is equipped with a basic HTTP REST API to support its operation from a variety of applications and external services. The operation of the accompanying application (Figure 2) are used to build the diagram in Figure 1 and to obtain the values in Table 1. Since the whole slide SVS image is de-referenced from a URL, the full file can come from either a public image repository such as TCGA (The Cancer Genome Atlas) or from a private Box. The prototype application uses a TCGA hosted image as a default use-case, but the user is encouraged to point it to private accounts in Box.com and DropBox (Figure 2, see also videcast link in the project github page), as well as other SVS files.
in the public domain such as TCGA’s. As noted in the application user interface (see link under Box buttons), the approach can be extended to many more Box providers using their API directly, and through commercial proxy services such as filepicker.com.

Figure 1 - ImageBox implementation architecture. At the most basic level, by exposing an image traversing API (in red), the cloud hosted imageBox server acts as a cache for an IIP slice server. That is the case when the external whole slide SVS file is hosted in a public store such as TCGA’s (see Methods). The engagement of regular file-picker APIs of cloud stores such as Box.com and DropBox is the most novel aspect of this architecture, because it allows the serving of image slices without breaching the access constraints of the original image. A user can of course only push to imageBox the files she has access, and in response will receive the unique dereferenceable identifier of the private images being sliced. Finally, the slice URLs (see Figure 2) can be used by Web Apps and other client applications: see project github page for examples of direct use of imageBox from Matlab. This implementation architecture can therefore be straightforwardly extended (dashed lines) to image analysis services.
Figure 2 - Snapshot of accompanying ImageBox Web Application at http://sbu-bmi.github.io/imagebox, used to demonstrate interoperability features on the proposed slice distribution implementation architecture (Figure 1). A videocast of its usage is available in the project home page at https://github.com/SBU-BMI/imagebox.

The ImageBox application prototype was pointed to a variety of whole slide image SVS files in Box.com and Dropbox cloud folders, as well as in public folders such as The Cancer Genome Atlas (TCGA, see Methods for folder address). The ImageBox slice server (See Methods) was in turn hosted in Amazon AWS. Although transfer rates varied, we found that between the within-cloud resource components the values are consistently above 10 MBs, and are routinely twice that much (Table 1). The slice download rates, even when the client machine was in a network equipped with generous bandwidth, did not typically exceed 1.5 MBs. The more important number, however, is the availability of individual image slices resolving significant morphological features in under a second, and while placing no storage requirements on the client application. The support for slice URL’s (Fig.2, highlighted in yellow: click on “open in new page”) has two important attributes. The first one is that they are portable, readily dereferenceable, and can be analyzed and annotated by any client application without requiring the whole image. The second critical advantage is that they do not expose the original hosting source, as can be verified by noting that the same image is rendered with different URLs when retrieved from a private Box folder by two different users, or even by the same user in two different sessions. These values and behaviors can be verified by using the accompanying ImageBox prototype in Fig 2, or by visualizing the webcast on the project github page.
Table 1 - Example transfer rates from Dropbox to the imageBox in Fig 2 (http://Nitrous.IO). As is often the case, both resources use Amazon's AWS as their backend, which implies that benchmarking was in fact performed on 10Gbs connected AWS EC2 servers for both EBS-backed system volumes, as well as, SSD-backed system volumes. The reader is encouraged to generate their own measurements by using the accompanying imageBox application (see Availability).

<table>
<thead>
<tr>
<th>SVS file size (MB)</th>
<th>Time to transfer (seconds)</th>
<th>Data Transfer Rate (MB/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,064.96</td>
<td>38</td>
<td>28</td>
</tr>
<tr>
<td>449.4</td>
<td>16</td>
<td>28.1</td>
</tr>
<tr>
<td>171.06</td>
<td>5.3</td>
<td>32.3</td>
</tr>
<tr>
<td>175.49</td>
<td>7.1</td>
<td>24.7</td>
</tr>
<tr>
<td>116.3</td>
<td>3.4</td>
<td>34.2</td>
</tr>
<tr>
<td>75.15</td>
<td>2.3</td>
<td>32.7</td>
</tr>
<tr>
<td>2,052.36</td>
<td>72.1</td>
<td>28.5</td>
</tr>
</tbody>
</table>

As discussed in the next section, the scalability of the imageBox architecture benefits from the 10-fold faster within-cloud transfer rates. However, the scalability of the proposed implementation architecture does not: the faster rates only have an effect on the availability of the first slice for a given whole slide image. The slower “first slice” reflects its use as the event that triggers transfer of the full image to the in-cloud imageBox IIP server (Fig1, “imageBox”). The retrieval of subsequent slices benefits from its ready availability in the imageBox slide server (Fig1 “ImageBox”) for retrieval by a client application (Fig 1, “WebApps” and “Web Services”). It is really in these subsequent traversals of the pre-existing image that the promise of “cloudified” scalable image analysis is realised. Case in point, pathologically relevant image features, at any scale, should be resolvable by 1,000² and surely by 2,000² image slices. It is worth noting that the image analysis operations on such slices typically exceed the 1 and 4 secs, respectively, that it takes to retrieve them. In summary, with the imageBox implementation architecture, the slice retrieval ceases to be the limiting step, which then moves to the availability of distributed computational resources. Just as important, the nature and governance of those resources is now much broader than in the conventional architecture, for example when using OpenSlide, as will be discussed below.

Discussion

Given its well deserved popularity and wide use of OpenSlide [Goode 2013], this discussion of the imageBox implementation architecture has to start there. OpenSlide’s ensemble of an efficient C-based image server and a client side dynamic zoom library established and popularised the core features of availability and scalability of a slice server architecture in general. In that regard, the work described here lays on top of those achievements rather than
representing an alternative. The innovative element advanced by imageBox is the programmatic intermediation for distributed image analysis instead of visualization. Accordingly, the emphasis was placed on a) reducing the functionality of the slice server down to the bare minimum of responding to the selection of a scale and a pixel slice, b) wrapping the slice server as a cloud hosted HTTP REST service such that c) it could be programmatically operated not only for public image’s (such as TCGA’s), but also for images privately hosted in stores equipped with file-picker APIs (“Boxes”, see Introduction).

The novel operational features of the imageBox deployment architecture are particularly clear when considering a data-driven, consumer-facing approach. As the accompanying prototype demonstrates, the user of this Web Application does not have to download or maintain any server side resource. Correspondingly, the App developer does not have access, or maintenance responsibilities, to the slice server or the storage where the data is primarily hosted. In a nutshell, the imageBox implementation architecture was designed for full decoupling between application development and application usage. Although, along the lines of QMachine [Wilkinson 2014], one could distribute in-browser image analysis using approaches such as imageJS [Almeida 2012], a more immediate choice would be to expose established image analysis applications, such as 3D Slicer [Fedorov 2012], as web services. That approach, however, passes image sections to a remote location, which may raise privacy concerns or have logistical dependencies the domain user and third party client applications cannot control.

For the reasons discussed above, the prototype deployment accompanying this report stops with the demonstration that a web application (client) is able to retrieve image slices from whole slide images hosted in a private store (making use of the file-picker/Box API) without placing any requests that the user deploys or logs into a server side resource. Furthermore, the inspection of the slice URL composition will reveal a basic HTTP REST API which can be engaged from computational environments well suited for image analysis, not only 3D Slicer [Fedorov 2012] as noted above, but also general purpose image analysis environments such as Matlab (see project github repository for examples). Given the successful delivery of a functional prototype for SVS whole slide images, we would argue that one or more cloud based imageBox slice servers would be a valuable addition as public resource for researchers developing or consuming image analysis applications. We envision that such a resource would be further developed to support a wider range of image formats, and might be configured to accept only a preset list of origins for the full image files. In any case, by removing the need to download the full images and engaging the corresponding specialized analytical libraries [Teodoro 2013] [Ali 2013] programmatically, the imageBox architecture opens the distribution of domain facing applications to any Biomedical Informatics Group able to develop them. The critical data-driven nature of this architecture is that the application code can be executed by those with access to the data, without compromise of the image privacy or any requirement to support server side components. The advantages of the resulting consumer-facing architectures go beyond the promise of commoditizing specialized image analysis operations - they are at the very core of the emergence of new data-intensive disciplines studying disease [Colen 2014] such as Computational Pathology [Louis 2015, Roth 2015]. More to the point, the imageBox mediated
analysis of large images such as whole slides, is reduced to engaging the mediation of a HTTP
REST application programming interface (API) of a cloud computing resource.

Conclusion

The imageBox implementation architecture was designed to address user-facing large image
analysis scenarios. To demonstrate the feasibility of a data-driven deployment, a web
application was developed to call whole slide images in SVS format from both public (TCGA)
and private (Box.com, DropBox) stores. This open source application is made freely available at
s bu-bmi.github.io/imagebox both to demonstrate the feasibility of the imageBox architecture,
and to offer reference for other box deployments that may emerge to balance the load of their
usage. This report argues that the cost-effective distribution achieved in this manner identifies
novel opportunities for cloud-based image analysis along a path pioneered by other domains of
application such as computational genomics.

Acknowledgements

This work was supported in part by 1U24CA180924-01A1 from the NCI, R01LM011119-01 and
R01LM009239 from the NLM.

References

Data Warehousing System over MapReduce. Proceedings VLDB Endowment. 2013
Almeida JS, E Iriabho, VL Gorrepati, S Wilkinson, DE Robbins, A Grüneberg, JR Hackney
(2012) ImageJS: personalized, participated, pervasive and reproducible image
Box, Box for Healthcare, retrieved Sep 2015 from www.box.com/healthcare.
Report: Clinical and Computational Requirements for Correlating Imaging Phenotypes with
Genomics Signatures. Transl Oncol. 2014 Oct 24;7(5):556-69. doi:
25389451; PubMed Central PMCID: PMC4225695.
[PMID: 24244884].
Google Inc, HIPAA Compliance & Data Protection with Google Apps - HIPAA implementation
guide (2015)
1/hipaa_implementation_guide.pdf.


An Extended SNOMED CT Concept Model for Observations in Molecular Genetics

James R. Campbell MD1, Geoffrey Talmon MD1, Allison Cushman-Vokoun MD PhD1, Daniel Karlsson PhD2, W. Scott Campbell PhD1
1University of Nebraska Medical Center, Omaha NE USA; 2Department of Biomedical Engineering, Linköping University; Linköping, Sweden

Abstract

Molecular genetics laboratory reports are multiplying and increasingly of clinical importance in diagnosis and treatment of cancer, infectious disease and managing of public health. Little of this data is structured or maintained in the EHR in format useful for decision support or research. Structured, computable reporting is limited by non-availability of a domain ontology for these data. The IHTSDO and Regenstrief Institute(RI) have been collaborating since 2008 to develop a unified concept model and ontology of observable entities – concepts which represent the results of laboratory and clinical observations. In this paper we report the progress we have made to apply that unified concept model to the structured recording of observations in clinical molecular genetic pathology including immunohistochemistry and sequence variant findings. The primary use case for deployment is the structured and coded reporting of Cancer checklist© and biomarker data as developed by the College of American Pathologists(CAP) with collaboration by the Royal College of Pathology(RCP).

Introduction

Molecular genetic pathology is a new scientific frontier exploding on the practice of clinical medicine. President Obama pushed the issue to the fore when he announced the national agenda for research into personalized medicine1. Unfortunately the extensive work in developing and managing information in genetic research has not translated into ontologies of use in support and documentation of clinical practice. Of the reference terminologies cited by the Office of the National Coordinator (ONC) as required by the US healthcare information architecture2, only LOINC3 has significant content addressing observables in molecular genetics. Unfortunately the LOINC reference terminology model applied to molecular genetics does not capture important details such as laboratory methods4 or support needs of the domain ontology that would support re-use of observational data in epidemiology, research and clinical decision making5. Diverse efforts within the informatics community to develop a clinically useful observables ontology6-11 have been informative but none have gained broad acceptance for integration into the ONC terminology architecture for the electronic health record (EHR).

Laboring behind the scenes in worldwide terminology management, harmonization efforts by the International Health terminology Standards Development Organization (IHTSDO) and Regenstrief Institute (RI)12 have been quietly working to expand the expressivity and utility of observable entities and clinical findings for use in molecular genetic structured clinical data. An observable entity is a concept with semantic overlap between LOINC and SNOMED CT (<<363787002|Observable entity|) and can best be described as a conceptual model for the results of an observation – administrative, clinical, laboratory or otherwise. Although RI has served the informatics community for years providing LOINC codes for laboratory medicine and molecular genetic pathology, the lack of a domain ontology for these concepts means that the there is no terminology support for queries of aggregation or for defining features such as methods4,5. The harmonization work underway has developed a candidate unified concept model for genetic observables but the application of that work is not intuitive. This paper reports one small part of that effort focused upon the challenging issues of genetic biomarker observations in anatomic pathology supporting diagnosis and management of cancer.

The volume and types of molecular genetic data appearing in clinical medicine is growing exponentially. CAP first published cancer report protocols (checklists)13 in 1998 defining a minimum dataset for anatomic pathologists to
report when they evaluate surgical specimens concluding a diagnosis of cancer. Currently CAP publishes 82 separate checklists for various tissue pathways. The checklists were expanded in 2013 to include tumor biomarkers which have become referent findings required for staging and planning treatment.

The number of these genetic observations – regarding either the tumor or the patient – important to outcomes in cancer treatment has grown with the science. An example we encountered while modeling the colorectal cancer check sheet was “detection of the BRAF V600E sequence variant in the resected tumor”. This genetic finding has been studied and clinically validated to predict response to certain treatments in both colorectal cancer and melanoma. Historically the clinical pathologist could only detect genetic sequence variations by analyzing extracted tumor DNA for specific mutations in single genes that code for the mutant proteins they formed. Those tests yielded limited analysis of genetic material and were expensive and time consuming. With the successful mapping of the human genome and improved technology, high throughput sequencing of human and neoplastic genomes became possible. Next generation sequencing (NGS) yields relatively rapid turn-around of much more genetic sequence data. Anatomic pathology today employs both protein and nucleotide sequence data in the diagnosis, prognostication and selection of targeted therapeutic regimens for cancer.

Primary use cases for a domain ontology of observables are query support of findings aggregated by genotypic variant, tissue of origin and histologic appearance. Query use cases implied by this expectation might include: “Find all cases of malignancies that tested positive for the BRAF p.V600E (c.1799T>A) mutation”; “Find all patients who have tested positive for genetic predisposition to breast cancer”; “Find all cases of colon cancer in which no biomarker testing was performed”

The SNOMED CT concept model\textsuperscript{14} consists of a constrained set of relationships and accompanying target value sets allowed for incorporation within a computable concept definition. Allowed relationships and value sets are specified for domains of SNOMED CT, usually individual hierarchies. A SNOMED CT concept must also have at least one supertype (IS_A) relationship linking the concept within the hierarchy and a fully specified (context free) name which is the universal term denoting the concept. Each linguistic implementation of SNOMED CT can have one primary term and as many synonyms as required. When the concept model applied to the modelled SNOMED CT content is insufficient to fully define a concept, the concept is declared as Primitive. The SNOMED CT ontology is subjected to description logic classification prior to publication as an editorial quality check and to compute the inferred relationships implied by the application of the concept model. The concept model is not complete for all segments of SNOMED CT. For years the Observable entities hierarchy (\texttt{<<363787002|Observable entity|}) has been published as a hierarchy of Primitives employing only stated supertype relationships.

The IHTSDO expects that extensions of SNOMED CT may be authored, compliant with the concept model, that represent material necessary for parochial or research needs not appropriate for the international release of SNOMED CT. In the US, the National Library of Medicine (NLM) develops and maintains the US extension to SNOMED that is required for Meaningful Use compliance by EHR vendors. The University of Nebraska maintains the Nebraska Lexicon\textsuperscript{©} extension, dependent upon the US extension, which supports terminology needs of our Epic\textsuperscript{®} implementation and terminology development we have been supporting for the community since 2004. We report in this paper our use of this terminology authoring environment to develop, test and deploy an observables ontology for molecular genetics requirements of the CAP cancer check sheets.

Concept model development

The Observables and Investigation Model Project was formed by the IHTSDO in 2008 to develop a computable concept model for the Observable entity domain and prepare a model for interoperation of content with LOINC data sets. In summer of 2015, the project convened a meeting of experts from the NLM, CAP and Health and Social Care Information Center (HSCIC) of UK to discuss details of application of the proposed concept model to the set of observables necessary to the structured reporting of cancer checklists and biomarkers. During the meeting, data elements contained in the CAP colorectal check sheet and the Royal College of Pathologists (RCP) counterpart were reviewed and analyzed for semantics. Pathologists provided expert direction regarding the clinical meaning of each
element. Terminology experts then discussed the requirements and proposed templates for consistent application of a candidate concept model for observables. The deployment model for testing from that conference is pictured in Figure 1 for an Observable entity in anatomic and molecular pathology. Each attribute for refining the meaning of a concept is shown along with the valuesets of target concepts that are supported as well as the associated cardinality of the relationship. Non-defining relationships, called qualifiers, are pictured in brown in this figure.

![Figure 1. Harmonized observable concept model](image)

Application of this model was reasonably straightforward for conventional observations in surgical pathology, requiring only model extensions of SNOMED to include 8 Properties and 4 Techniques. The Colon cancer checklist which we first modeled required 61 new observable entities for anatomic pathology.

Consensus on application of this model to observables in molecular genetic pathology could not be developed in our first three meetings and a deployment strategy was only achieved at the Montevideo meeting of the IHTSDO in fall 2015. Criticisms emerging from initial discussions included: a) excessive numbers of primitives, b) insufficient semantic granularity in genetic structures and c) failure to support both protein-based and sequence-based observations. In order to meet these challenges we proposed employing Human Genome Nomenclature Committee\textsuperscript{15} resource data in our model by map reference to uniquely define nucleotide sequence entities (genes, microsatellites and other nucleotide segments) and to support the details of sequence data observations in Variant Call Format\textsuperscript{16}. We also employed representation of numbers in our model for nucleotide references\textsuperscript{17}. We proposed that a gene locus and other sequence based data could be defined in an expanded concept model specifying the gene locus as cellular substructures, specifically nucleotide sequences. These data were further characterized using the reference naming of the Human Gene Nomenclature Committee (HGNC)\textsuperscript{15} and the genetic datasets which it cross references such as the annotated genome reference Ensembl\textsuperscript{19}. An example concept model rendition for the B-RAF proto-oncogene as we defined it is included in figure 2. This particular concept definition employs sequence address
data from the Genome Reference Consortium GRCh38 release in GenBank\textsuperscript{20}, included in our publication by reference as a map data set. The map data which is shown in tabular form as an inset includes the HGNC reference number and a REST service call which will retrieve full HGNC reference data. Human protein products of gene transcription are currently implemented as primitive concepts in the 105590001|Substance| hierarchy.

**Figure 2.** Cellular structure model for BRAF gene locus

**Figure 3.** Immunohistochemistry observable: BRAF protein expression by immunoperoxidase staining
Once the reference genetic material was fully defined in SNOMED, we extended Techniques and Properties for molecular genetic pathology procedures. Immunohistochemistry observations could then be modelled. An example of an immunoperoxidase staining analysis for BRAF protein in a surgical tumor specimen is shown in figure 3. Immunoperoxidase and nuclear sequencing techniques are much more specific than the molecular procedures currently employed in LOINC 2.54 concept definitions. Therefore the observables concepts we have modeled for CAP checklists are generally semantic children of molecular genetic observables now in LOINC.

Nucleotide sequence observables were a particularly thorny issue we faced since input from our pathologist specialists required that we be able to store observations in our research databases with complete sequence data. An industry standard for identifying sequencing results has become the Variant Call Format\textsuperscript{16}. These data files issue findings of sequence variants when compared to a reference standard genome and are typically ASCII files of a few kilobytes. The report data structures complies with the recommendations for description of sequence variants issued by the Human Genome Variant Society\textsuperscript{18}. Once again, employing numbers as SNOMED CT values allowed us to extend findings data and to deploy Has value attributes in place of Has interpretation in an extended concept model for Clinical findings. Figures 4 and 5 show the Observable entity model for sequence data of the BRAF gene locus along with a positive clinical finding for the BRAF p. V600E (c.1799T>A) mutation detected in the excised tumor.
Application of the harmonized concept model and modelling of the SNOMED CT extension concepts to the recording needs of the check sheets was the first step in preparing a completely specified structured report for CAP cancer synoptics. For each observable entity representing a ‘question’ on the form, we further organized SNOMED CT value sets of ‘answers’ that would populate the attribute-value pair of the HL7 OBX segment. For the example of “BRAF protein expression by immunoperoxidase staining in excised malignant neoplasm” presented in figure 3, we show the corresponding section of the CAP check sheet for colorectal cancer in figure 6. Included in the figure is an inset with the Observable entity concept identifier, fully specified name and the complete valueset of choices that should be selected by the pathologist in the anatomic pathology system.

Figure 6. CAP check sheet for colorectal cancer; biomarkers segment for BRAF genetic observations

Results

Our deployment use cases are the 82 cancer and tumor biomarkers checklists published by CAP and supplemented by review with the RCP. There are many repeating observations across these checklists but our experience with deployment of our model for colon cancer required expansion of the Machine Readable Concept Model21 for the SNOMED CT hierarchies of Observables, Body structure and Clinical Findings. The magnitude of the new conceptual content in our Nebraska Lexicon© extension required for colon and breast cancer is summarized in table 1. This lists the number of new concepts by hierarchy which we developed for anatomic and molecular genetic pathology. The number of primitive concepts we required for each domain is also listed. In the fourth column we have listed an exemplar concept name from the content developed within that SNOMED CT hierarchy. This work has proceeded in collaboration with NLM, RI, IHTSDO, CAP and the RCP and was published on the NLM UMLS Knowledge Sources Server for public dissemination in July 2016. The work is available to all interested parties with a UMLS license.
Table 1. Extension concept inventory for colon cancer checklist

<table>
<thead>
<tr>
<th>SNOMED CT hierarchy</th>
<th>Anatomic Pathology Concepts/Primitives</th>
<th>Molecular Genetic Concepts/Primitives</th>
<th>Exemplar molecular extension concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable entities</td>
<td>61/1</td>
<td>32/3</td>
<td>BRAF nucleotide sequence detected in excised malignancy</td>
</tr>
<tr>
<td>Body Structures</td>
<td>10/9</td>
<td>29/3</td>
<td>BRAF gene locus</td>
</tr>
<tr>
<td>Clinical findings</td>
<td>6/2</td>
<td>7/3</td>
<td>BRAF V600E variant identified in excised malignancy</td>
</tr>
<tr>
<td>Procedures</td>
<td>2/1</td>
<td>0</td>
<td>Pyrosequencing</td>
</tr>
<tr>
<td>Techniques</td>
<td>4/4</td>
<td>7/7</td>
<td>Sequence property</td>
</tr>
<tr>
<td>Property types</td>
<td>8/8</td>
<td>2/2</td>
<td>Variant call format</td>
</tr>
<tr>
<td>Scale types</td>
<td>0</td>
<td>9/9</td>
<td></td>
</tr>
<tr>
<td>Situations</td>
<td>1/0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Substances</td>
<td>0/0</td>
<td>11/11</td>
<td>BRAF human cellular protein</td>
</tr>
<tr>
<td>Attributes</td>
<td>2/2</td>
<td>3/3</td>
<td></td>
</tr>
<tr>
<td>Qualifiers</td>
<td>2/2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>88/29</td>
<td>100/41</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

As predicted, sequencing of the human genome has led to a proliferation of innovative scientific research with application to clinical medicine. We expect that new types of clinical knowledge will create challenges for recording, managing and using the clinical data that results from that research. While a number of efforts have been reported by the informatics community\textsuperscript{3-11} to organize or develop terminology models to serve for recording of clinical genomic data, only LOINC offers substantial content within the suite of ONC US standards. Unfortunately, the LOINC model does not offer the features of a domain ontology that are desirable for best query and retrieval of clinical data.

We have presented in summary form the results of a two year collaboration with clinical and terminology standards developers which represents only a new snapshot of harmonization work between the IHTSDO and RI in process since 2008. This report focuses on Observable entities for molecular genetic observations because they represent new challenges to the application of the SNOMED CT/LOINC harmonized concept model. Such advances are a challenging but important component of a comprehensive terminology model for twenty-first century medicine. CAP cancer check sheets offer one set of important use cases for clinical ontology development because they summarize anatomic and molecular genetic observations of established relevance for practice of precision medicine.

We have deployed this model for testing and evaluation by other informatics centers and published the work in collaboration with the NLM. We expect that this will stimulate dialogue with the informatics terminology community on the ontology model we have deployed. We hope to scale the work across the rest of molecular genetic pathology observations for cancer and expand the work into microbiology and human germline genetic disease testing.

The model we present interacts and draws upon NCBI resources and ontologies supported by the Genome Reference Consortium\textsuperscript{22}. We expect that the science in this field will continue to rapidly evolve and that new concepts and observations will emerge from that research requiring documentation in the EHR. The role of ONC terminologies and, more specifically clinical ontologies, should be to faithfully record those data and provide for query and re-use for purposes of clinical decision support, research and public health. We do not however, think it the role of SNOMED CT or LOINC to secondarily reproduce GRC reference data. For that reason, HGNC defining data sets are included in the model by map reference.

358
References

Analyzing SNOMED CT’s Historical Data: Pitfalls and Possibilities

Werner Ceusters, MD¹, Jonathan P. Bona, PhD¹
¹Department of Biomedical Informatics, University at Buffalo, Buffalo, NY

Abstract
SNOMED CT’s Release Format 2 (RF2) has been announced as an improvement over its predecessor, for instance because of its more consistent and almost formal approach towards describing changes in components over different versions, as well as changes in the structure of SNOMED CT itself. We explore two sorts of changes that are only partially formalized in RF2: the relationships between associative relations and reasons for inactivations as expressed in Association Reference Sets and Attribute Value Reference Sets on the one hand, and the various patterns according to which semantic tags appearing in fully specified names change over subsequent versions with or without being related to inactivations. We propose a data conversion methodology that combines assertions about SNOMED CT components into history profiles and use elements of these profiles to build Formal Concept Analysis contexts to discover valid implications that can render implicit assumptions hidden in SNOMED CT’s structure explicit.

Introduction
SNOMED CT [1], maintained by the International Health Terminology Standards Development Organization (IHTSDO) is a large healthcare terminology built around a concept-based ontology. Concepts are classified under several hierarchies, of which most of the top classes correspond roughly to the types of entities instances of which are encountered by clinicians during their work (body parts, diseases, substances, procedures, etc.) while other top classes correspond to types instantiated by descriptive elements of the SNOMED CT knowledge representation itself, for example classes denoted by terms such as ‘inactive concept’, ‘navigational concept’, and ‘core metadata concept’ [2]. In addition to active components – ‘component’ being the umbrella term used by IHTSDO for concept or relationship or description – SNOMED CT contains also inactive components which were active in one or more prior versions but at some point have been inactivated for one or other reason. Prior to July 2011 all releases were distributed in a format now known as ‘Release Format 1 (RF1)’. Release Format 2 (RF2) was introduced in 2012 in order to (1) implement a more robust and consistent representation of versions in which changes are tracked in a uniform manner in the core files themselves; (2) introduce reference sets as a more easily extensible and maintainable replacement for the RF1 representations of subsets for specific uses such as not only mappings to other biomedical terminologies and classification systems, but also to improve RF1’s history mechanism; and (3) create an added hierarchy to represent metadata about the structure of SNOMED CT itself [3, 4 p127].

The part of SNOMED CT that describes its own history has grown considerably over the years, as witnessed, for example, by the 105,313 inactive concepts – roughly 25% of the total concept count – which are annotated by 140,390 associations to other concepts or descriptions. For 99,489 of these inactive concepts reasons for their inactivations are provided. This comes on top of the 325,893 reasons for description-inactivations related to both active and inactive concepts. This raises the question whether the totality of assertions which are about changes in SNOMED CT rather than about external reality constitutes in and of itself a valuable resource to identify patterns that would allow the detection of mistakes in assertions about external reality that have thus far not been discovered. In other words: what can we learn about SNOMED CT’s mistakes committed in the past to detect still existing mistakes and prevent new ones? This sort of quality improvement being the ultimate goal of our efforts, the work described in this paper is the first phase of this endeavor during which we explored the history information included in the RF2 distribution of the January 2016 version of SNOMED CT to identify pitfalls and possibilities that should be taken into account for the development of novel error detection methods.

Changes in SNOMED CT
The content of SNOMED CT evolves with each release. Once released, SNOMED CT components are persistent and their identifiers are not reused [4, p45]. When a component becomes inactive this is indicated by the value of the active field, a field which is present in all components. Components continue to be distributed even when they are no longer active. This allows a current release to be used to interpret data entered using an earlier release. Within RF2, all changes in components are represented in the corresponding files by adding a new row, with the same component ID, a new effective time and any necessary change in the component values. As an example, Table 1 shows that the concept ‘301381004’ with FSN ‘Discomforting present pain (finding)’ was set to active in release 20020131 and to inactive in 20080131.
Table 1. Updates in the SNOMED CT concept file (RF2) for concept 301381004 with FSN ‘Discomforting present pain (finding)’.  

<table>
<thead>
<tr>
<th>conceptID</th>
<th>Effective Time</th>
<th>Active</th>
<th>ModuleID</th>
<th>Definitional Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>301381004</td>
<td>20020131</td>
<td>1</td>
<td>9000000000000207008</td>
<td>90000000000000074008</td>
</tr>
<tr>
<td>301381004</td>
<td>20080131</td>
<td>0</td>
<td>9000000000000207008</td>
<td>90000000000000074008</td>
</tr>
</tbody>
</table>

Legend: Active: (1) = active, (0) = inactive.

Table 2. Updates in the SNOMED CT relationships file (RF2) for the same concept 301381004

<table>
<thead>
<tr>
<th>RelID</th>
<th>Effective Time</th>
<th>Active</th>
<th>Attribute</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>126300024</td>
<td>20020131</td>
<td>1</td>
<td>Is a</td>
<td>Pain (finding)</td>
</tr>
<tr>
<td>126300024</td>
<td>20040131</td>
<td>0</td>
<td>Is a</td>
<td>Pain (finding)</td>
</tr>
<tr>
<td>126301023</td>
<td>20020131</td>
<td>1</td>
<td>Is a</td>
<td>Finding of present pain intensity (finding)</td>
</tr>
<tr>
<td>126301023</td>
<td>20080131</td>
<td>0</td>
<td>Is a</td>
<td>Finding of present pain intensity (finding)</td>
</tr>
<tr>
<td>657858027</td>
<td>20020131</td>
<td>1</td>
<td>Finding site</td>
<td>Structure of nervous system (body structure)</td>
</tr>
<tr>
<td>657858027</td>
<td>20060131</td>
<td>0</td>
<td>Finding site</td>
<td>Structure of nervous system (body structure)</td>
</tr>
<tr>
<td>2260209021</td>
<td>20030731</td>
<td>1</td>
<td>Interprets</td>
<td>Nervous system function (observable entity)</td>
</tr>
<tr>
<td>2260209021</td>
<td>20050131</td>
<td>0</td>
<td>Interprets</td>
<td>Nervous system function (observable entity)</td>
</tr>
<tr>
<td>2458913020</td>
<td>20040131</td>
<td>1</td>
<td>Is a</td>
<td>Discomfort (finding)</td>
</tr>
<tr>
<td>2458913020</td>
<td>20080131</td>
<td>0</td>
<td>Is a</td>
<td>Discomfort (finding)</td>
</tr>
<tr>
<td>2858465020</td>
<td>20060131</td>
<td>1</td>
<td>Finding site</td>
<td>Anatomical structure (body structure)</td>
</tr>
<tr>
<td>2858465020</td>
<td>20080131</td>
<td>0</td>
<td>Finding site</td>
<td>Anatomical structure (body structure)</td>
</tr>
</tbody>
</table>

Legend: RelID = Relationship identifier; Active: (1) = active, (0) = inactive. Columns irrelevant for our purposes here are not shown. For readability, Attribute and Target identifiers have been replaced by their corresponding FSN – omitting ‘(attribute)’ – in the most recent version studied (January 2016).

Table 2 shows that during the life time of that concept, it underwent considerable changes in its reported relationships to other concepts after full DL classification. Table 3 demonstrates how changes in the descriptions of concepts are similarly logged. Only one description record with the same descriptionID field is current at any point in time. The current record is the one with the most recent Effective Time before or equal to the point in time under consideration. If the active field is false (‘0’), then the description is inactive at that point in time. If it is true (‘1’), then the description is associated with the concept identified by the conceptId field (not shown in Table 3).

Table 3. Updates in the SNOMED CT descriptions file (RF2) for concept ‘274236006’

<table>
<thead>
<tr>
<th>descriptionID</th>
<th>Effective Time</th>
<th>Active</th>
<th>Description Type</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>410015012</td>
<td>20020131</td>
<td>1</td>
<td>Synonym</td>
<td>Asthenia</td>
</tr>
<tr>
<td>410015012</td>
<td>20040131</td>
<td>0</td>
<td>Synonym</td>
<td>Asthenia</td>
</tr>
<tr>
<td>666971011</td>
<td>20020131</td>
<td>1</td>
<td>FSN</td>
<td>Asthenia [D]</td>
</tr>
<tr>
<td>666971011</td>
<td>20030131</td>
<td>0</td>
<td>FSN</td>
<td>Asthenia [D] (finding)</td>
</tr>
<tr>
<td>1237162017</td>
<td>20020731</td>
<td>1</td>
<td>Synonym</td>
<td>Asthenia [D]</td>
</tr>
<tr>
<td>1472277017</td>
<td>20030131</td>
<td>1</td>
<td>FSN</td>
<td>Asthenia (context-dependent category)</td>
</tr>
<tr>
<td>1472277017</td>
<td>20060731</td>
<td>0</td>
<td>FSN</td>
<td>Asthenia (context-dependent category)</td>
</tr>
<tr>
<td>1489933012</td>
<td>20030131</td>
<td>1</td>
<td>Synonym</td>
<td>Asthenia</td>
</tr>
<tr>
<td>2610401019</td>
<td>20060731</td>
<td>1</td>
<td>FSN</td>
<td>Asthenia (situation)</td>
</tr>
</tbody>
</table>

Legend: Active: (1) = active, (0) = inactive. Columns irrelevant for our purposes here are not shown. For readability, Description Type identifiers have been replaced by their corresponding term – omitting their semantic tag ‘(core metadata concept)’.

RF2 replaces the ‘history mechanism’ implemented in RF1 [5] by means of Historical Association Reference Sets (HARS) and Component Inactivation Reference Sets (CIRS). HARSs (Table 4) are used to indicate, for example, which deactivated concepts are in one way or another related to other active concepts, and CIRSs (Table 5) to indicate the reasons for inactivating a component – such as errors, duplication of another component, and ambiguity of meaning [4, p506]. Records that express such association are called reference set members. The primary purpose of these
reference sets is to specify which (if any) of these associations should be followed in a fashion similar to following ‘Is a (attribute)’ relations when determining whether to retrieve a record entry previously coded with a concept that has since then been inactivated. Whereas ‘same as’ and ‘replaced by’ associations can be followed unproblematically, the solution for ambiguous concepts related by ‘possibly equivalent to’ associations is less clear-cut [4, p654].

**Table 4.** Historical association reference set types in SNOMED CT (modified from [4, p509])

<table>
<thead>
<tr>
<th>HARS name</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possibly equivalent to (P)</td>
<td>From an ambiguous concept to one or more active concepts that represents one of the possible meanings of the inactive concept.</td>
</tr>
<tr>
<td>Moved to (T)</td>
<td>From a component to a namespace to which the component has been moved</td>
</tr>
<tr>
<td>Moved from (F)</td>
<td>From a namespace to the original component Identifier in its previous namespace.</td>
</tr>
<tr>
<td>Replaced by (R)</td>
<td>From an erroneous or obsolete inactive component to a single active replacement component.</td>
</tr>
<tr>
<td>Same as (S)</td>
<td>From a duplicate component to the active component that this component duplicates.</td>
</tr>
<tr>
<td>Was a (W)</td>
<td>From an inactive classification concept such as &quot;not otherwise specified&quot; to the active concept that was formerly its most proximal supertype.</td>
</tr>
<tr>
<td>Alternative (Z)</td>
<td>From an inactive classification concept derived from ICD-9 Chapter XVI 'Symptoms signs and ill-defined conditions' with the most similar active concept.</td>
</tr>
<tr>
<td>Refers to</td>
<td>From an inactive description which is inappropriate to the concept it is directly linked to but instead should refer to the concept referenced.</td>
</tr>
</tbody>
</table>

**Table 5.** Component inactivation set types for concepts (modified from [4, p506-507])

<table>
<thead>
<tr>
<th>CIRS value</th>
<th>Concept status and motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicate (D)</td>
<td>inactive because it has the same meaning as another Concept</td>
</tr>
<tr>
<td>Outdated (O)</td>
<td>inactive because it is an outdated concept that is no longer used.</td>
</tr>
<tr>
<td>Ambiguous (A)</td>
<td>inactive because it is inherently ambiguous either because of an incomplete FSN or because it has several associated terms that are not regarded as synonymous or partial synonymous.</td>
</tr>
<tr>
<td>Erroneous (E)</td>
<td>inactive because it contains an error</td>
</tr>
<tr>
<td>Limited (L)</td>
<td>active prior to Jan 2010, inactive since then because of unstable meaning within SNOMED CT</td>
</tr>
<tr>
<td>moved to (M)</td>
<td>inactive because moved to another namespace.</td>
</tr>
<tr>
<td>Pending move</td>
<td>active but in the process of being moved to another namespace.</td>
</tr>
</tbody>
</table>

**Methodology**

SNOMED CT undergoes changes of various sorts with each release. Most changes are recorded explicitly in the sense that there is a formal mechanism through which changes of this type are documented in one or other component of the SNOMED CT distribution files. Some types of changes, implicit ones, lack such a formal mechanism but can be retrieved through the implementation of algorithms not documented in the SNOMED CT documentation. In this paper, we report on the evolution of activations and inactivations of component instances as examples of an explicit type of change, and on the evolution of semantic tags as an implicit type of change. Both types of changes required specific data reorganization strategies. The results of these conversions were then analyzed using Formal Concept Analysis.

**Data reorganization of activations and inactivations**

We combined the explicit assertions about (in)activations in components and history information present in CIRSS and HARSs into one format. We kept track of in which versions assertions were made – and possibly also changed – through the construction of a history profile. Such profile contains for each of the 29 versions from January 2002 to January 2016 a marker indicating whether the assertion was in that version absent (A) or present, in which case it was either active (Y) or inactivated (N). Since the very same concepts can not only appear as referenced component in one HARS member and as target component in another HARS member, but also appear in members of distinct HARSs, it was possible to compute clusters of concepts by randomly selecting a concept from a HARS member and recursively collecting all reference set members in which this concept appears with the goal of processing each associated concept in the same way until no more concepts can be found. Table 6 contains assertions that were retrieved for one such cluster composed out of 5 related concepts. By ‘assertion’, we mean anything that inside SNOMED CT is explicitly
or implicitly stated as applying to a concept. Examples of explicit assertions are the inclusion of the concept 324253001 in the 1st version (‘Y’ in first position of the history profile in row 8), the inactivation of that concept in the 3rd version (‘N’ in 3rd position, row 8) and the reactivation of it in the 5th version.

Table 6. History profile of SNOMED CT assertions related to a cluster of 5 concepts about Azithromycin dihydrate

<table>
<thead>
<tr>
<th>Row</th>
<th>ConceptID</th>
<th>Attribute</th>
<th>Value</th>
<th>History profile (one character per version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>324253001</td>
<td>Duplicate</td>
<td></td>
<td>AAYYNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Duplicated by</td>
<td>375558000</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Duplicated by</td>
<td>375559008</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Duplicated by</td>
<td>375948007</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Duplicated by</td>
<td>376025007</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Semantic tag</td>
<td>product</td>
<td>VNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Semantic tag</td>
<td>substance</td>
<td>YNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Is active</td>
<td></td>
<td>YNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Same-as</td>
<td>375559008</td>
<td>AAYYNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Same-as</td>
<td>375948007</td>
<td>AAYYNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
</tr>
<tr>
<td>11</td>
<td>375558000</td>
<td>Duplicate</td>
<td></td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Semantic tag</td>
<td>product</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Is active</td>
<td></td>
<td>AYNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>Same-as</td>
<td>324253001</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Same-as</td>
<td>375948007</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>16</td>
<td>375559008</td>
<td>Duplicate</td>
<td></td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>Duplicated by</td>
<td>324253001</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>Is active</td>
<td></td>
<td>AYNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>Same-as</td>
<td>324253001</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>Same-as</td>
<td>375948007</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>21</td>
<td>375948007</td>
<td>Duplicate</td>
<td></td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>Duplicated by</td>
<td>324253001</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>Duplicated by</td>
<td>375558000</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>Duplicated by</td>
<td>375559008</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>Duplicated by</td>
<td>376025007</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>Semantic tag</td>
<td>Product</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>Is active</td>
<td></td>
<td>AYNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>Same-as</td>
<td>375948007</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>29</td>
<td>376025007</td>
<td>Duplicate</td>
<td></td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>Semantic tag</td>
<td>Product</td>
<td>AAAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td>Is active</td>
<td></td>
<td>AYNNNNNNNNNYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>Same-as</td>
<td>324253001</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>33</td>
<td></td>
<td>Same-as</td>
<td>375948007</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
</tbody>
</table>

Legend: concepts in cluster with most recent FSN: 324253001: Azithromycin 200mg/5mL oral suspension (product), 375558000: Azithromycin dihydrate 200mg/5mL suspension (product), 375559008: Azithromycin dihydrate 200mg/5 mL suspension (product), 376025007: Azithromycin dihydrate 200mg/5 mL powder (product). History profile: ‘A’: absent, ‘Y’: active, ‘N’: inactive.

These assertions are explicit because there are corresponding effective time and activation status assertions in the concept table. Similarly, there is in the description table the assertion that the semantic tag of this concept was in the 1st version ‘substance’ (row 7), while ‘product’ in the 2nd version (row 6). Examples of implicit assertions are, for instance, the ones to the effect that in the 1st version there was no mention of a sameness-relation between this concept and 375559008 (row 9) nor 375948007 (row 10).

Data reorganization for the evolution of semantic tags

We computed for each concept the evolution of what in SNOMED CT is called semantic tags. Descriptions provide for each concept a Fully Specified Name (FSN) most of which ‘end with a semantic tag in parentheses and which
indicates the semantic category to which the concept belongs (e.g. clinical finding, disorder, procedure, organism, person, etc.)’ [4, p41]. It is further stated that ‘the semantic tag helps to disambiguate different concepts which may be referred to by the same commonly used word or phrase’ [4, p41]. For example, it is the semantic tag ‘morphologic abnormality’ in the FSN ‘Hematoma (morphologic abnormality)’ that disambiguates the concept to which this FSN is assigned from a second concept with FSN ‘Hematoma (disorder)’. The former is intended to be used for what ‘a pathologist sees at the tissue level’, while the latter ‘represents the clinical diagnosis that a clinician makes when they decide that a person has a “hematoma”’ [4, p41]. As can already be seen in Table 6 (rows 6 and 7) concepts can be assigned different semantic tags over time, changes which for some, as exemplified by Table 7 are quite dramatic.

Table 7. Examples of changes in semantic tag assignment over time

<table>
<thead>
<tr>
<th>conceptID</th>
<th>Most recent FSN</th>
<th>Changes in semantic tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>66076007</td>
<td>Chewable tablet (qualifier value)</td>
<td>(substance)</td>
</tr>
<tr>
<td>66402002</td>
<td>Peritoneal dialysis education (procedure)</td>
<td>(procedure)</td>
</tr>
<tr>
<td>68433009</td>
<td>Childhood (finding)</td>
<td>(function)</td>
</tr>
<tr>
<td>69736008</td>
<td>Vocational assessment (procedure)</td>
<td>(procedure)</td>
</tr>
<tr>
<td>70409003</td>
<td>Mouthwash (qualifier value)</td>
<td>(substance)</td>
</tr>
<tr>
<td>70444001</td>
<td>Recessive gene (substance)</td>
<td>(function)</td>
</tr>
<tr>
<td>70790008</td>
<td>Absence of nausea and vomiting (situation)</td>
<td>(finding)</td>
</tr>
<tr>
<td>73669007</td>
<td>Kung fu (qualifier value)</td>
<td>(qualified value)</td>
</tr>
<tr>
<td>73905001</td>
<td>Sees flickering lights (finding)</td>
<td>(qualified value)</td>
</tr>
</tbody>
</table>

Legend: semantic tags are written between brackets. ‘|’ indicates a transition from one (or more) tags to another. History profiles are omitted in this table.

Data analysis

Exploratory statistical analyses were performed to find associations, or unexpected lack thereof, between the various sorts of assertions derived from the conversions. Specifically to mention here is Formal Concept Analysis (FCA), a mathematical theory for understanding the structure of data given as a set of objects described in terms of attributes they possess, which is done by representing the data as a concept lattice [6]. Every FCA concept – we will use explicitly the term ‘FCA concept’ to distinguish it from SNOMED CT concepts – has its extent (the set of objects that fall under the FCA concept) and its intent (the set of attributes that together are necessary and sufficient for an object to be an instance of the FCA concept. In [7], for instance, attributes were defined on the basis of the normal forms of pre-coordinated SNOMED CT expressions. For our analyses here, we created FCA attributes and corresponding objects on the basis of two contexts: (1) the co-occurrence of SNOMED CT HARS and CIRS attributes as described in Table 4 and Table 5 throughout the history of SNOMED CT concepts, and (2) the evolution of semantic tags over time. While FCA concept lattice diagrams have visualizing power when applied to domains with a small number of concepts and attributes governed by a simple organizational structure, they are rather useless in case of more complex situations as the one explored here. More useful here is the computation of attribute implications where an implication asserts a certain relationship between two attribute sets which are respectively called premise and conclusion: an implication is valid in the data set if every object that has all attributes from the premise of the implication also has all attributes from its conclusion. For example, if ‘being mammal’ and ‘being vertebrate’ would be attributes used to correctly describe animals, then ‘being mammal $\rightarrow$ being vertebrate’ would be computable as being a valid implication. The set of all valid implications can be reduced to a smaller set – the Duquenne–Guigues base – from which all other implications follow semantically [8], and a set of approximate implications known as Luxenburger base [6]. The latter are valid for a specified percentage of FCA concepts; for example, for a particular zoo it could be found that 85% of the vertebrates on display are mammals. For both contexts, we assessed whether implications correspond to SNOMED CT’s editorial policies in relation to inactivations and SNOMED CT’s concept model.

Results and discussion

Concept ambiguity and possibly-equivalent-to associations

SNOMED CT’s technical implementation guide [4] contains indications that certain changes in components go hand in hand with changes in HARSs and CIRSs. For instance, from the description of what it means for an inactive concept
to be *possibly equivalent to* another concept (Table 4) one can assume that such concepts are asserted as being ambiguous in a CIRSs. We found that not to be the case for 4 concepts. On the other hand, although it is allowed for an inactivé concept to be *possibly equivalent to* only one other concept – we found 7815 of such cases – it is for many of these cases hard, if not impossible, to find out, especially algorithmically through some automated procedure, why the change has been made in this specific way. It is clear that ‘Pyogenic arthritis of lower leg (disorder)’ is ambiguous in the sense that it does not specify in which joint specifically the arthritis is located, but then the question is why it has only been associated with ‘Knee pyogenic arthritis (disorder)’. Another example is ‘Distal interphalangeal joint structure of third finger (body structure)’ which has been asserted as being *possibly equivalent to* another concept with exactly the same FSN. Inactivation because of ambiguity is stated to be ‘because it is inherently ambiguous either because of an incomplete FSN or because it has several associated terms that are not regarded as synonymous or partial synonymous’ (Table 5). Since there was nothing wrong with the FSN it must have been because of the synonyms. Indeed, inspection reveals that for the inactivated concept there is the synonym ‘Distal interphalangeal joint of third digit of forelimb’ which is not to be found in the target concept, thus removing the ambiguity of whether the concept denotes a human body part in addition to an animal body part. This reasoning, unfortunately, does not hold for ‘391651001: Gluten-free/wheat-free baguette (product)’ which was rendered ambiguous in the 5th release and made *possibly equivalent to* 407775004, a then newly introduced concept with exactly the same set of terms (Table 8). Relevant questions are (1) what motivated the SNOMED CT editor to introduce the new concept, and (2) why to use the inactivation because of ambiguity rather than because of duplication? The history mechanism is not able to provide arguments.

**Table 8. Inactivation of ‘391651001: Gluten-free/wheat-free baguette (product)’**

<table>
<thead>
<tr>
<th>ConceptID</th>
<th>Attribute</th>
<th>Value</th>
<th>History profile (one character per version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>391651001</td>
<td>AMB</td>
<td></td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td></td>
<td>Is active</td>
<td></td>
<td>AAYYNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
</tr>
<tr>
<td></td>
<td>Poss-equivalent-to</td>
<td>407775004</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td>407775004</td>
<td>Semantic tag</td>
<td>product</td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
<tr>
<td></td>
<td>Is active</td>
<td></td>
<td>AAAAYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY</td>
</tr>
</tbody>
</table>

**Duplication and same-as associations**

**Table 4 and Table 5** let us reasonably assume that when a duplicate concept is discovered, it is inactivated with the reason for inactivation being stated as ‘duplicate’ and that at the same time a *same-as* association would be created. That is indeed the case for 44,113 inactivations. Yet, we found 1449 cases where a concept is stated to be duplicate without a corresponding same-as association, and 3453 cases in which a same-as association was created without a duplicate assertion.

We also found cases in which concepts are stated to be duplicates, yet denote clearly distinct entities. The most notorious case is the one in which ‘34759008: Urethral catheter, device (physical object)’ is stated to be duplicated by 73 other concepts, each one of which denoting nevertheless a more precisely specified type of catheter, for example ‘349499005: Bard 10mL balloon 22Ch 1658 2-way all-silicone male length urethral Foley catheter’ and ‘349501002: Bard 10mL balloon 24Ch 1265LV 2-way Teflon coated male urethral Foley catheter’. Several of these catheters are nevertheless, by means of other concepts, listed as descendants of ‘34759008’.

**Mining implication and association rules for concept inactivations**

The observation that – with one exception: whenever a ‘moved to’ association is asserted in a HARS, there is a corresponding ‘moved to’ inactivation asserted in a CIRS – there are no simple consistent relationships between reasons for inactivation and associations motivated us to mine for possible relationships using Formal Concept Analysis. In our case here, we used as FCA concepts every combination of CIRS and HARS membership encountered for all (active and inactive) SNOMED CT concepts as depicted in Table 6, without, however, including history profile information. We used the uppercase characters written between brackets in the first columns of Table 4 and Table 5 – with the exception of ‘pending move’ and ‘refers-to’ which were not included in this analysis – to name our FCA concepts. For example, the FCA concept ‘DLSW’ was attached to every SNOMED CT concept for which throughout its history it was at least once stated to be duplicate (D) and limited (L), as well as associated with other concepts by means of a same-as (S) and was-a (W) association. 85 such FCA concepts were found, in total covering all 424,759 active and inactive SNOMED CT concepts.
Table 9 shows the positive Duquenne-Guigues and Luxenburger base of implications between HARS and CIRS assertions. For example, the implication ‘< 3 > ADW ==» S’ states that if a SNOMED CT concept has ever been annotated as being ambiguous, duplicate and enjoying a was-a association to some other SNOMED CT concept, then it is also the case that this concept has been annotated as having a same-as association. The ‘<3>’ indicates that this implication corresponds to 3 of the 85 FCA concepts encountered which when calculated back to SNOMED CT concepts covers only 4 cases. As another example, the implication ‘< 5 > P E ==» A R’ states that whenever a SNOMED CT concept has been asserted to be possibly equivalent to some other concept as well as being erroneous, it is also the case that that SNOMED CT concept has been annotated as being ambiguous and having been replaced by some other concept. This is the case for 5 FCA concepts which cover in total 15 SNOMED concepts.

The Luxenburger base implication ‘< 24 > P ==» A’ states that for 92% of the FCA concepts which correspond to SNOMED CT concepts that have been annotated with a possibly equivalent to association it is the case that the corresponding SNOMED CT concepts have also been annotated as being ambiguous. Similarly, ‘< 20 > E ==» R’ states that for 90% of the FCA concepts with an E-attribute, there is also an R attribute. When these two implications are assessed towards the SNOMED CT concepts to which the FCA concepts apply, then we find 18,413 SNOMED CT concepts under the 22 FCA concepts for which the P ==» A implication holds, and 1,397 SNOMED CT concepts under the 18 FCA concepts for which E ==» R holds. These counts are many magnitudes higher than the 15 SNOMED CT concepts for which P E ==» A R holds. This raises the question whether this huge discrepancy is indicative for something being wrong with all or some of the assertions made in relation to these 15 concepts. And it leads to the more general question whether the differences in counts observed between SNOMED CT concepts covered by Duquenne–Guigues implications in contrast to Luxenburger implications form the basis for a novel method of quality control that to the best of our knowledge has thus far not been applied to SNOMED CT.

Table 9. Partial Duquenne-Guigues and Luxenburger base of implications between HARS and CIRS assertions

<table>
<thead>
<tr>
<th>Positive implications from the Duquenne–Guigues base</th>
<th>Luxenburger base &gt;80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;19&gt;M==»T; &lt;3&gt;ADW==»S; &lt;1&gt;MT E==»R; &lt;24&gt;P[92%]==»&lt;22&gt;A;</td>
<td></td>
</tr>
<tr>
<td>&lt;19&gt;T==»M; &lt;1&gt;R OW==»L; &lt;5&gt;A E==»P R; &lt;20&gt;E[90%]==»&lt;18&gt;R;</td>
<td></td>
</tr>
<tr>
<td>&lt;9&gt;P S==»A; &lt;2&gt;Z==»M T; &lt;5&gt;P E==»A R; &lt;10&gt;P R[90%]==»&lt;9&gt;A;</td>
<td></td>
</tr>
<tr>
<td>&lt;4&gt;A R S==»P; &lt;1&gt;M T A L==»P; &lt;7&gt;S E==»R; &lt;10&gt;A R[90%]==»&lt;9&gt;P;</td>
<td></td>
</tr>
<tr>
<td>&lt;1&gt;M T A D==»P; &lt;8&gt;P L==»A; &lt;2&gt;A P R S E==»D; &lt;9&gt;A L[89%]==»&lt;8&gt;P;</td>
<td></td>
</tr>
<tr>
<td>&lt;3&gt;A R D==»P S; &lt;2&gt;R S L==»D; &lt;8&gt;D E==»R; &lt;9&gt;P D[89%]==»&lt;8&gt;A;</td>
<td></td>
</tr>
<tr>
<td>&lt;3&gt;P R D==»A S; &lt;9&gt;D L==»S; &lt;2&gt;O E==»R; &lt;7&gt;W E[86%]==»&lt;6&gt;R;</td>
<td></td>
</tr>
<tr>
<td>&lt;2&gt;P O==»A; &lt;1&gt;R O L==»W; &lt;1&gt;R S O E==»D; &lt;7&gt;R S E[86%]==»&lt;6&gt;D;</td>
<td></td>
</tr>
<tr>
<td>&lt;1&gt;A R O==»P; &lt;1&gt;O W L==»R; &lt;5&gt;L E==»R; &lt;6&gt;A W[83%]==»&lt;5&gt;L;</td>
<td></td>
</tr>
<tr>
<td>&lt;3&gt;D O==»S; &lt;1&gt;S F==»D; &lt;2&gt;A R W==»P E; &lt;6&gt;A W[83%]==»&lt;5&gt;P;</td>
<td></td>
</tr>
<tr>
<td>&lt;2&gt;M T W==»S L; &lt;1&gt;D F==»S; &lt;2&gt;R S W==»D E; &lt;11&gt;A S[82%]==»&lt;9&gt;P;</td>
<td></td>
</tr>
<tr>
<td>&lt;5&gt;P W==»A; &lt;2&gt;A R L==»P E; &lt;3&gt;R D W==»E; &lt;5&gt;A W L[80%]==»&lt;4&gt;P;</td>
<td></td>
</tr>
</tbody>
</table>

Semantic tag evolutions

We found in total 285 patterns of the sort exemplified in Table 7 according to which SNOMED CT concepts underwent changes in the semantic tags assigned to them. A change from no semantic tag at all to a semantic tag (43 patterns) counted – under one perspective – also as a change. There were no patterns with more than 3 changes over time under either perspective. Changes in semantic tags can happen for a number of reasons. One is a change in SNOMED CT’s concept model, for instance when distinctions are made that didn’t exist in earlier versions, or different interpretations were introduced (e.g. the product / substance distinction). Such changes have a global impact on large parts of the ontology. Another reason is that concepts were in one or other way erroneous and had to be corrected. The SNOMED CT documentation states for instance that ‘only limited changes may be made to the “term” field, as defined by editorial rules’ [4, p145]. This is consistent with the view that ‘the meaning of a concept can be determined [...] from associated descriptions that include human readable terms’ [4, p87]. This editorial rule is also used as argument for not retiring the concept to which it is attached in cases where the FSN undergoes minor changes. Indeed, ‘Minor changes in the FSN are those changes that do not alter its meaning. A change to the semantic type shown in parentheses at the end of the FSN may sometimes be considered a minor change if it occurs within a single top-level hierarchy (e.g. a change from a finding tag to a disorder tag, or a change from a procedure tag to a regime/therapy tag), but a move to a completely different top-level hierarchy is regarded as a significant change to
the Concept’s meaning is prohibited" [4, p393]. It was therefore hypothesized that changes in semantic tags within the history of a concept would strongly correlate with the concept being inactive in the most recent version. A first analysis summarized in the right half of Table 10 under the perspective of ‘no tag → semantic tag’ constituting a change shows that this is indeed the case for concepts in which only one such change occurred: where the expected ratio of active versus inactive concepts is 75.2%/24.8%, the observed ratio is 28.5%/71.5%. The Cramer’s V statistic over the entire table being 0.55, suggests a strong correlation. That there are more than expected active concepts with more than one change might be explained by the corrections of mistakes made during an earlier change. However, as further inspection revealed that the majority of one-time changes were changes from no semantic tag to a semantic tag, we recalculated the matrix under this 2nd perspective (left half of Table 10) only to find out that under this perspective no conclusive association is present (Cramer V=0.09). It was also at that time that our attention was drawn to the fact that a large amount of semantic tags were assigned to the FSN of concepts that were already inactive since many earlier versions!

Table 10. Associations between number of semantic tag changes and inactivations

<table>
<thead>
<tr>
<th>Frequency Overall %</th>
<th>No tag → semantic tag = no change</th>
<th></th>
<th></th>
<th>No tag → semantic tag = change</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Active Observed</td>
<td>Inactive Expected</td>
<td>Active Inactive</td>
<td>Active Observed</td>
<td>Inactive Expected</td>
<td>Active Inactive</td>
</tr>
<tr>
<td></td>
<td>1 0</td>
<td>1 0</td>
<td>1 0</td>
<td>1 0</td>
<td>1 0</td>
<td>1 0</td>
</tr>
<tr>
<td>0</td>
<td>292,823</td>
<td>90,798</td>
<td>383,621</td>
<td>288,508</td>
<td>95,113</td>
<td>334,719</td>
</tr>
<tr>
<td></td>
<td>68.9% 21.4% 90.3%</td>
<td>67.9% 22.4%</td>
<td>75.2% 24.8%</td>
<td>68.9% 9.9% 78.8%</td>
<td>87.4% 12.6%</td>
<td>75.2% 24.8%</td>
</tr>
<tr>
<td></td>
<td>76.3% 23.7%</td>
<td>70.3% 90.3%</td>
<td>90.3% 75.2%</td>
<td>91.6% 39.9%</td>
<td>78.8% 78.8%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>25,027</td>
<td>14,403</td>
<td>39,430</td>
<td>29,654</td>
<td>9,776</td>
<td>39,430</td>
</tr>
<tr>
<td></td>
<td>5.9% 34.9% 9.3%</td>
<td>7.0% 23.7%</td>
<td>75.2% 24.8%</td>
<td>5.9% 14.9% 20.8%</td>
<td>28.5% 71.5%</td>
<td>75.2% 24.8%</td>
</tr>
<tr>
<td>2</td>
<td>1,543</td>
<td>107</td>
<td>1,650</td>
<td>1,241</td>
<td>409</td>
<td>1,241</td>
</tr>
<tr>
<td></td>
<td>0.4% 0.0% 0.4%</td>
<td>0.3% 0.1%</td>
<td>75.2% 24.8%</td>
<td>0.4% 0.0% 0.4%</td>
<td>0.5% 6.9%</td>
<td>0.3% 1.1%</td>
</tr>
<tr>
<td></td>
<td>93.5% 6.5%</td>
<td>0.5% 0.1%</td>
<td>93.1% 0.5%</td>
<td>93.1% 6.9%</td>
<td>0.5% 0.1%</td>
<td>0.3% 1.1%</td>
</tr>
<tr>
<td>3</td>
<td>53</td>
<td>5</td>
<td>58</td>
<td>44</td>
<td>14</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>0.0% 0.0% 0.0%</td>
<td>0.0% 0.0%</td>
<td>75.2% 24.8%</td>
<td>0.0% 0.0% 0.0%</td>
<td>0.0% 0.0%</td>
<td>0.0% 0.0%</td>
</tr>
<tr>
<td></td>
<td>91.4% 8.6%</td>
<td>0.0% 0.0%</td>
<td>91.4% 8.6%</td>
<td>0.0% 0.0% 0.0%</td>
<td>0.0% 0.0%</td>
<td>0.0% 0.0%</td>
</tr>
<tr>
<td>Totals</td>
<td>319,446</td>
<td>105,313</td>
<td>424,759</td>
<td>334,719</td>
<td>105,313</td>
<td>440,052</td>
</tr>
</tbody>
</table>

A second observation was that certain change patterns occur frequently within a smaller subset of semantic tags. One such subset is the one formed by the semantic tags disorder, finding, situation, morphologic abnormality, event and navigational concept. We constructed again a formal concept analysis context on the basis of 13 attributes: 2 in relation to each of the semantic tags in the subset just sketched, each such attribute reflecting whether the tag is one which occurs in a non-terminal position or a terminal position in a change pattern, and one reflecting whether the concept is active in the last version. We computed the number of SNOMED CT concepts that are described by means of this FCA context (Table 11, not showing, however, the breakdown in active/inactive concepts). This table shows, for instance, that – within this context as specified – what finally became tagged as events, where primarily tagged as findings in some earlier version, as well as to a large extent disorders, with the exception of 25 concepts that started with a semantic tag outside the subset of 6.

We also computed the corresponding Duquenne–Guigues base with the number of SNOMED CT concepts that are covered by these implications (Table 12). Although this cluster applies to 20,867 concepts, there are not many cases that are covered by the positive implications in the Duquenne–Guigues base. The implication with the largest number of FCA concepts (i.e. 4) s → A covers only 40 SNOMED CT concepts. It tells us that – within this context – all SNOMED CT concepts that had a situation in non-terminal position are active. From Table 11 we can read that 14 of
these concepts were finally tagged as findings, and 26 as navigational concepts. The low coverage of positive implications suggests that hard rigor in semantic tag change patterns is hard to come by.

**Table 11.** Distribution of semantic tag change patterns within a subset of 6 semantic tags

<table>
<thead>
<tr>
<th>Non-terminal tag</th>
<th>Terminal tag: disorder</th>
<th>situation</th>
<th>Morphologic abnormality</th>
<th>finding</th>
<th>event</th>
<th>Navigational Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>D</td>
<td>S</td>
<td>M</td>
<td>F</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>s</td>
<td>278</td>
<td>141</td>
<td>0</td>
<td>299</td>
<td>1303</td>
<td>185</td>
</tr>
<tr>
<td>m</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1650</td>
<td>0</td>
<td>472</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>8124</td>
<td>215</td>
</tr>
<tr>
<td>e</td>
<td>271</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>n</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o</td>
<td>71</td>
<td>4464</td>
<td>99</td>
<td>2949</td>
<td>25</td>
<td>268</td>
</tr>
<tr>
<td></td>
<td>Total concepts: 2279</td>
<td>5077</td>
<td>100</td>
<td>3263</td>
<td>9452</td>
<td>696</td>
</tr>
</tbody>
</table>

**Table 12.** Positive implications from the Duquenne–Guigues base of an FCA context built out the semantic tags specified in Table 11

<table>
<thead>
<tr>
<th>Implication</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 4 &gt; s ====&gt; A</td>
<td>40</td>
</tr>
<tr>
<td>&lt; 1 &gt; e N ====&gt; A d</td>
<td>2</td>
</tr>
<tr>
<td>&lt; 1 &gt; A m ====&gt; M</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 1 &gt; A s f ====&gt; N</td>
<td>6</td>
</tr>
<tr>
<td>&lt; 1 &gt; m M ====&gt; A</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 4 &gt; d f ====&gt; A</td>
<td>20</td>
</tr>
<tr>
<td>&lt; 1 &gt; A d s ====&gt; F</td>
<td>5</td>
</tr>
<tr>
<td>&lt; 1 &gt; f F ====&gt; A</td>
<td>1</td>
</tr>
</tbody>
</table>

**Limitation: what qualifies as semantic tags?**

The SNOMED CT documentation available from the IHTSDO webserver provides insufficient information on what the precise set of semantic tags the SNOMED CT editors are working with might be. The information that a semantic tag is that what appears at the end of a FSN between brackets [4, p41] is not reliable. Historically, FSNs didn’t have a semantic tag at all, as this was apparently introduced later as witnessed by the many changes in descriptions to that end. Parsing anything that terminates a FSN between brackets leads to many false positives in older concepts. For many of those, manual inspection is required for disambiguation. But even then it is not always obvious especially in light of the occurrence of FSNs that apparently enjoy 2 semantic tags: Figure 1 depicts all terms which we assume to be (or have been at some point) semantic tags and what they are collocated with. For example, we found 393 FSN assertions in which the semantic tag ‘body structure’ collocates with other semantic tags, i.e. ‘morphologic abnormality’ (11 occurrences), ‘surface region’ (82 occurrences) and ‘combined site’ (300 cases).

**Conclusion**

SNOMED CT has undoubtedly come a very long way since its original conception as a mere nomenclature for pathology [9, 10]. The IHTSDO has been working very hard on developing editorial and technical principles for updating SNOMED CT and on training its terminologists in applying the principles faithfully. Furthermore, the distribution format RF2 presents itself as a formidable resource to obtain a deeper insight in how SNOMED CT evolved. The exploratory analyses we have performed as part of the work described here and which are of sorts that to our best knowledge have thus far not been described in the literature, made us aware of certain possibilities, but nevertheless revealed many pitfalls in attempting to derive from SNOMED CT’s history mechanism what the before mentioned principles exactly might be, or whether they are indeed applied consistently. Whether it is the methodology proposed here itself, or a lack of, for instance, discriminatory power in the reasons for inactivation – one could even wonder why no reasons are given for the addition of new concepts –, is something that needs further to be researched. Nevertheless, it is at this stage of our work possible to formulate the following concrete recommendations towards the IHTSDO: (1) formalize the relationships between semantic tags and SNOMED CT concept hierarchies, (2) implement in the authoring environment mechanisms to prevent and detect incoherent and missing CIRS and HARS records, and (3) provide reasons for not only inactivations, but also activations, which reflect whether changes are purely internal in SNOMED CT (e.g. because of changes in the concept model) or external (changes in the covered domains).
Figure 1. Co-occurrence of candidate semantic tags in FSNs. Numbers inside nodes tally the FSNs in which the candidate semantic tag in the node collocates with another tag. Numbers along the edges tally the FSNs in which the tags in the connected nodes collocate with each other.

Acknowledgments

This work was supported in part by Clinical and Translational Science Award NIH 1 UL1 TR001412-01 from the National Institutes of Health, and by grant R21LM009824 from the National Library of Medicine (NLM). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCR, the NLM or the National Institutes of Health.

References

Interpretable Deep Models for ICU Outcome Prediction

Zhengping Che, Sanjay Purushotham, PhD, Robinder Khemani, MD, Yan Liu, PhD
1University of Southern California, Los Angeles, CA, USA
2Children’s Hospital Los Angeles, Los Angeles, CA, USA

Abstract
Exponential surge in health care data, such as longitudinal data from electronic health records (EHR), sensor data from intensive care unit (ICU), etc., is providing new opportunities to discover meaningful data-driven characteristics and patterns of diseases. Recently, deep learning models have been employed for many computational phenotyping and healthcare prediction tasks to achieve state-of-the-art performance. However, deep models lack interpretability which is crucial for wide adoption in medical research and clinical decision-making. In this paper, we introduce a simple yet powerful knowledge-distillation approach called interpretable mimic learning, which uses gradient boosting trees to learn interpretable models and at the same time achieves strong prediction performance as deep learning models. Experiment results on Pediatric ICU dataset for acute lung injury (ALI) show that our proposed method not only outperforms state-of-the-art approaches for morality and ventilator free days prediction tasks but can also provide interpretable models to clinicians.

1 Introduction
The national push for electronic health records (EHR) has resulted in an exponential surge in volume, detail, and availability of digital health data. This offers an unprecedented opportunity to infer richer, data-driven descriptions of health and illness. Clinicians are collaborating with computer scientists to improve the state of health care services towards the goal of Personalized Healthcare. Unlike other data sources, medical/hospital data such as EHR is inherently noisy, irregularly sampled (or have missing value), and heterogeneous (data come from different sources such as lab tests, doctor's notes, monitor readings etc). These data properties make it very challenging for most existing machine learning models to discover meaningful representations or to make robust predictions. This has resulted in development of novel and sophisticated machine learning solutions. Among these methods, deep learning models (e.g., multilayer neural networks) have achieved the state-of-the-art performance on several tasks, such as computational phenotype discovery and predictive modeling.

Even though powerful, deep learning models (usually with millions of model parameters) are difficult to interpret. In today's hospitals, model interpretability is not only important but also necessary, since clinicians are increasingly relying on data-driven solutions for patient monitoring and decision-making. An interpretable predictive model is shown to result in faster adoptability among clinical staff and better quality of patient care. Decision trees, due to their ease of interpretation, have been successfully employed in the health care domain, and clinicians have embraced them for predictive tasks such as disease diagnosis. However, decision trees can easily overfit and perform poorly on large heterogeneous EHR datasets. Thus, an important question naturally arises: how can we develop novel data-driven solutions which can achieve state-of-the-art performance as deep learning models and at the same time can be easily interpreted by health care professionals and medical practitioners?

Recently, machine learning researchers have conducted preliminary work aiming to interpret the learned features from deep models. An early work investigated visualizing the hierarchical representations learned by deep networks, while a followup work explored feature generalizibility in convolutional neural networks. More recent work argued that interpreting individual units of deep models can be misleading. This line of work has shown that interpreting deep learning features is possible but the behavior of deep models may be more complex than previously believed, which motivates us to find alternative strategies to interpreting how deep model work.

In the meanwhile, recent work showed empirically that shallow neural networks are capable of achieving similar prediction performance as deep neural networks by first training a state-of-the-art deep model, and then training a shallow neural networks using predictions by the deep model as target labels. Similarly, Hinton et. al proposed an efficient knowledge distillation approach to transfer (dark) knowledge from model ensembles into a single model following the idea of model compression. Another work takes a Bayesian approach to distill knowledge from a deep neural network to a shallow neural network. Furthermore, mimic learning has also been successfully applied...
to multitask learning, reinforcement learning and speech processing applications\textsuperscript{25,26,27}. These work motivate us to explore the possibility of employing mimic learning to learn an interpretable model and at the same time achieves similar performance as a deep neural network.

In this paper, we introduce a simple yet effective knowledge-distillation approach called interpretable mimic learning, to learn interpretable models with robust prediction performance as deep learning models. Unlike standard mimic learning\textsuperscript{21}, which uses shallow neural networks or kernel methods, our interpretable mimic learning framework uses gradient boosting trees (GBT)\textsuperscript{28} to learn interpretable models from deep learning models. GBT, as an ensemble of decision trees, provides good interpretability along with strong learning capacity. We conduct extensive experiments on several deep learning architectures including feed-forward networks\textsuperscript{29} and recurrent neural networks\textsuperscript{30} for mortality and ventilator free days prediction tasks on Pediatric ICU dataset. We demonstrate that deep learning approaches achieve state-of-the-art performance compared to several machine learning methods. Moreover, we show that our interpretable mimic learning framework can maintain strong prediction performance of deep models and provide interpretable features and decision rules.

2 Background and Deep Models

In this section, we will first introduce notations and describe two state-of-the-art deep learning models, namely feed-forward neural networks and gated recurrent unit. We use these two models (and their extensions) as baselines in our experiments as well as components of the proposed interpretable mimic learning.

2.1 Notations

EHR data from ICU contains both static variables such as general descriptors (demographic information collected during admission) and temporal variables, which possibly come from different modalities, such as injury markers, ventilator settings, blood gas values, etc. We use $X$ to represent all the input variables, and a binary label $y \in \{0, 1\}$ to represent the prediction task outcome such as ICU mortality or Ventilator free days (VFD). We also use $x_t$ to denote the temporal variables observed at time $t$. Our goal is to learn an effective and interpretable function $F()$ which can be used to predict the value of $y$ given the input $X$.

2.2 Deep Learning Models

Feedforward Networks A multilayer feedforward network\textsuperscript{29} (DNN) is a neural network with multiple nonlinear layers and possibly one prediction layer on the top to solve classification task. The first layer takes the concatenation of static and flattened temporal variables as the input $X$, and the output from each layer is used as the input to the next layer. The transformation of each layer $l$ can be written as

$$X^{(l+1)} = f^{(l)}(X^{(l)}) = s^{(l)}(W^{(l)}X^{(l)} + b^{(l)})$$

where $W^{(l)}$ and $b^{(l)}$ are respectively the weight matrix and bias vector of layer $l$, and $s^{(l)}$ is a nonlinear activation function, which usually takes one of logistic sigmoid, tanh, or ReLU\textsuperscript{31}. For a feed-forward network with $L$ layers shown in Figure 1(a), the output of the top-most layer $y_{\text{fin}} = X^{(L)}$ is the prediction score, which lies in $[0, 1]$. People also usually treat the output of second top layer $X^{(L-1)}$ as the features extracted by DNN, and these features are usually helpful as inputs for other prediction models. We show the structure of DNN model in Figure 1(a). During training, we optimize the cross-entropy prediction loss between the prediction output and the true label.

Gated Recurrent Unit Recurrent neural network (RNN) models, such as Long Short Term Memory (LSTM)\textsuperscript{32} and Gated Recurrent Unit (GRU)\textsuperscript{33}, have been shown to be successful at handling complex sequence inputs and capturing long term dependencies. In this paper, we use GRU to model temporal modalities since it has a simpler architecture compared to classical LSTM and has been shown to achieve the state-of-the-art performance among all RNN models for modeling sequential data\textsuperscript{30}. The structure of GRU is shown in Figure 1(b). Let $x_t \in \mathbb{R}^P$ denotes the variables at time $t$, where $1 \leq t \leq T$. At each time $t$, GRU has a reset gate $r_t$ and an update gate $z_t$ for each of the hidden state $h_t$. The update function of GRU is shown as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1} + b)) \quad h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

where matrices $W_z, W_r, W, U_z, U_r, U$ and vectors $b_z, b_r, b$ are model parameters. At time $t$, we take the hidden
states $h_t$ and treat it as the output of GRU $x_{t+1}$ at that time. As shown in Figure 1(c), we flatten the output of GRU at each time step and add another sigmoid layer on top of them to get the prediction $y_{t+1}$.

**Combinations of deep models** One limitation of GRU is that it only aims to model temporal data, while usually both static and temporal features are available in EHR data from ICU. Therefore we propose a combination model of feed-forward network (DNN) and GRU. As shown in Figure 1(d), in the combination model, we use one DNN model to take static input features and one GRU model to take temporal input features. We then add one shared layer on top, which takes the features from both GRU and DNN to make prediction, and train all the parts jointly.

## 3 Interpretable Mimic Learning

In this section, we introduce the interpretable mimic learning method, which learns interpretable models and achieves similar performance as deep learning models. The proposed approach is motivated by recent development of deep learning in machine learning research and specifically designed for the health care domain.

### 3.1 Knowledge Distillation

The main idea of knowledge distillation is to first train a large, slow, but accurate model and transfer its knowledge to a much smaller, faster, yet still accurate model. It is also known as mimic learning, which uses a complex model (i.e., deep neural network, or an ensemble of network models) as a teacher/base model to train a student/mimic model (such as a shallow neural network or a single network model). The way of distilling knowledge, a.k.a. mimicking the complex models, is to utilize the soft labels learned from the teacher/base model as the target labels while training the student/mimic model. The soft label, in contrast to the hard label from the raw data, is the real value output of the teacher model, whose value usually ranges in [0, 1]. It is worth noting that a shallow neural network model is usually not as accurate as a deep neural network model, if trained directly on the same training data. However, with the help of the soft labels from deep models, the shallow model is capable of learning the knowledge extracted by the deep model and can achieve similar or better performance.

The reasons that the mimic learning approach works well can be explained as follows: Some potential noise and error in the training data (input features or labels) may affect the training efficacy of simple models. The teacher model may eliminate some of these errors, thus making learning easier for the student model. Soft labels from the teacher model are usually more informative than the original hard label (i.e. 0/1 in classification tasks), which further improves the student model. Moreover, the mimic approach can also be treated as an implicit way of regularization on the teacher model, whose value usually ranges in [0, 1]. It is worth noting that a shallow neural network model is usually not as accurate as a deep neural network model, if trained directly on the same training data. However, with the help of the soft labels from deep models, the shallow model is capable of learning the knowledge extracted by the deep model and can achieve similar or better performance.

The reasons that the mimic learning approach works well can be explained as follows: Some potential noise and error in the training data (input features or labels) may affect the training efficacy of simple models. The teacher model may eliminate some of these errors, thus making learning easier for the student model. Soft labels from the teacher model are usually more informative than the original hard label (i.e. 0/1 in classification tasks), which further improves the student model. Moreover, the mimic approach can also be treated as an implicit way of regularization on the teacher model, which makes the student model robust and prevents it from overfitting. The parameters of the student model can be estimated by minimizing the squared loss between the soft labels from the teacher model and the predictions by the student model. That is, given a set of data $\{X_i\}$ where $i = 1, 2, \cdots, N$ as well as the soft label $y_{s,i}$ from the teacher model, we estimate the student model $F(X)$ by minimizing $\sum_{i=1}^{N} \| y_{s,i} - F(X_i) \|^2$.

While existing work on mimic learning focus on model compression (via shallow neural networks or kernel methods), they cannot lead to more interpretable models, which is important and necessary in health care applications. To address this, we introduce a simple and effective knowledge-distillation approach called interpretable mimic learning, to learn interpretable models that mimic the performance of deep learning models. The main difference of our approach from
existing mimic learning approaches is that we use Gradient Boosting Trees (GBT) instead of another neural network as the student model since GBT satisfies our requirements for both learning capacity and interpretability. In the following sections, we describe GBT and our proposed interpretable mimic learning in more details.

3.2 Gradient Boosting Trees

Gradient boosting machines\(^{28,34}\) are a method which trains an ensemble of weak learners to optimize a differentiable loss function by stages. The basic idea is that the prediction function \(F(X)\) can be approximated by a linear combination of several functions (under some assumptions), and these functions can be sought using gradient descent approaches. Gradient Boosting Trees (GBT) takes a simple classification or regression tree as weak learner, and add one weak learner to the entire model per stage. At \(m\)-th stage, assume the current model is \(F_m(X)\), then the Gradient Boosting method tries to find a weak model \(h_m(X)\) to fit the gradient of the loss function with respect to \(F_m(X)\). The coefficient \(\gamma_m\) of the stage function is computed by the line search strategy to minimize the loss. To keep gradient boosting from overfitting, a regularization method called shrinkage is usually employed, which multiplies a small learning rate \(\nu\) to the stage function in each stage. The final model with \(M\) stages can be written as:

\[
F_M(X) = \sum_{i=1}^{M} \nu \gamma_i h_i(X) + \text{const}
\]

3.3 Interpretable Mimic Learning Framework

We present two general training pipelines within our interpretable mimic learning framework, which utilize the learned feature representations or the soft labels from deep learning models to help the student model. The main difference between these two pipelines is whether to take the soft labels directly from deep learning models or from a helper classifier trained on the features from deep networks.

In Pipeline 1 (Figure 2), we directly use the predicted soft labels from deep learning models. In the first step, we train a deep learning model, which can be a simple feedforward network or GRU, given the input \(X\) and the original target \(y\) (which is either 0 or 1 for binary classification). Then, for each input sample \(X\), we obtain the soft prediction score \(y_{nn} \in [0, 1]\) from the prediction layer of the neural network. Usually, the learned soft score \(y_{nn}\) is close but not exactly the same as the original binary label \(y\). In the second step, we train a mimic Gradient boosting model, given the raw input \(X\) and the soft label \(y_{nn}\) as the model input and target, respectively. We train the mimic model to minimize the mean squared error of the output \(y_m\) to the soft label \(y_{nn}\).

In Pipeline 2 (Figure 3), we take the learned features from deep learning models instead of the prediction scores, input them to a helper classifier, and mimic the performance based on the prediction scores from the helper classifier. For each input sample \(X\), we obtain the activations \(X_{nn}\) of the highest hidden layer, which can be \(X^{(L-1)}\) from an \(L\)-layer feed forward network, or the flattened output at all time steps from GRU. These obtained activations can be considered as the extracted representations from the neural network, and we can change the its dimension by varying the size of the neural networks. We then feed \(X_{nn}\) into a helper classifier (e.g., logistic regression or support vector machines), to predict the original task \(y\), and take the soft prediction score \(y_c\) from the classifier. Finally, we train a mimic Gradient boosting model given \(X\) and \(y_c\).

In both pipelines, we apply the mimic model trained in the last step to predict the labels of testing examples.
Our interpretable mimic learning approach has several advantages. First, our proposed approach can provide models with state-of-art prediction performance. The teacher deep learning model outperforms the traditional methods, and student gradient boosting tree model is good at maintaining the performance of the teacher model by mimicking its predictions. Second, our proposed approach yields more interpretable model than the original deep learning model, which is complex to interpret due to its complex network structures and the large amount of parameters. Our student gradient boosting tree model has better interpretability than original deep model since we can study each feature's impact on prediction and, we can also obtain simple decision rules from the tree structures. Furthermore, our mimic learning approach uses the soft targets from the teacher deep learning model to avoid overfitting to the original data. Thus, our student model has better generalizations than standard decision tree methods or other models, which tend to overfit to original data.

4 Experiments

We conduct experiments on a Pediatric ICU dataset to answer the following questions: (a) How does our proposed mimic learning framework perform when compared to the state-of-the-art deep learning methods and other machine learning methods? (b) How do we interpret the models learned through the proposed mimic learning framework? In the remainder of this section, we will describe the dataset, methods, empirical results and interpretations to answer the above questions.

4.1 Dataset and Experimental Design

We conduct experiments on a Pediatric ICU dataset collected at the Children's Hospital Los Angeles. This dataset consists of health records from 398 patients with acute lung injury in the Pediatric Intensive Care Unit at Children's Hospital Los Angeles. It contains a set of 27 static features such as demographic information and admission diagnoses, and another set of 21 temporal features (recorded daily) such as monitoring features and discretized scores made by experts, for the initial 4 days of mechanical ventilation. We apply simple imputation to fill in missing values, where we take the majority value for binary variables, and empirical mean for other variables. Our choice of imputation may not be the optimal one and finding better imputation methods is another important research direction beyond the scope of this paper. For fair comparison, we used the same imputed data for evaluation of all the methods.

We perform two binary classification (prediction) tasks on this dataset: (1) Mortality (MOR): we aim to predict whether the patient dies within 60 days after admission. 20.10% of all the patients are mortality positive (i.e., patients died). (2) Ventilator Free Days (VFD): we aim to evaluate a surrogate outcome of morbidity and mortality (Ventilator free Days, of which lower value is bad), by identifying patients who survive and are on a ventilator for longer than 14 days within 28 days after admission. Since here lower VFD is bad, it is a bad outcome if the value $\leq 14$, otherwise it is a good outcome. 59.05% of all the patients have VFD $> 14$.

4.2 Methods and Implementation Details

We categorize the methods in our experiments into the following groups:

- Baseline machine learning methods which are popular in healthcare domains: Linear Support Vector Machine (SVM), Logistic Regression (LR), Decision Trees (DT) and Gradient Boosting Trees (GBT).
- Deep network models: We use deep feed-forward neural network (DNN), GRU, and the combinations of them (DNN + GRU).
- Proposed mimic learning models: For each of the deep models shown above, we test both the mimic learning pipelines, and evaluate our mimic model (GBTmimic).

We train all the baseline methods with the same input, i.e., the concatenation of the static and flattened temporal features. The DNN implementations have two hidden layers and one prediction layer. We set the size of each hidden layer twice as large as input size. For GRU, we only use the temporal features as input. The size of other models are set to be in the same scale. We apply several strategies to avoid overfitting and train robust deep learning models: We train for 250 epochs with early stopping criterion based on the loss on validation dataset. We use stochastic gradient descent (SGD) for DNN and Adam with gradient clipping for other deep learning models. We also use weight regularizer and dropout for deep learning models. Similarly, for Gradient Boosting methods, we set the maximum number of boosting stages 100, with early stopping based on the AUROC score on validation dataset. We implement all baseline methods using the scikit-learn package and all deep networks in Theano and Keras platforms.
Table 1: Interpretable mimic learning classification results for two tasks. (mean ± 95% confidence interval)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MOR (Mortality)</th>
<th>VFD (Ventilator Free Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.6437 ± 0.024</td>
<td>0.7408 ± 0.054</td>
</tr>
<tr>
<td>LR</td>
<td>0.6615 ± 0.027</td>
<td>0.3736 ± 0.038</td>
</tr>
<tr>
<td>DT</td>
<td>0.6024 ± 0.013</td>
<td>0.4369 ± 0.016</td>
</tr>
<tr>
<td>GBT</td>
<td>0.7196 ± 0.028</td>
<td>0.4171 ± 0.040</td>
</tr>
<tr>
<td>Deep Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNN</td>
<td>0.7266 ± 0.089</td>
<td>0.4117 ± 0.122</td>
</tr>
<tr>
<td>GRU</td>
<td>0.7666 ± 0.063</td>
<td>0.4587 ± 0.104</td>
</tr>
<tr>
<td>DNN + GRU</td>
<td>0.7813 ± 0.028</td>
<td>0.4874 ± 0.051</td>
</tr>
<tr>
<td>Best Mimic Model</td>
<td>0.7898 ± 0.030</td>
<td>0.4766 ± 0.050</td>
</tr>
</tbody>
</table>

Table 2: Top features and their corresponding importance scores.

<table>
<thead>
<tr>
<th>Task</th>
<th>MOR (Mortality)</th>
<th>VFD (Ventilator Free Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>GBT</td>
<td>GBTmimic</td>
</tr>
<tr>
<td>Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PaO2-Day2</td>
<td>0.0339</td>
<td>MAP-Day1 (0.0423)</td>
</tr>
<tr>
<td>BE-Day0</td>
<td>0.0433</td>
<td>MAP-Day1 (0.0384)</td>
</tr>
<tr>
<td>MAP-Day1</td>
<td>0.0510</td>
<td>MAP-Day1 (0.0354)</td>
</tr>
<tr>
<td>PH-Day1</td>
<td>0.0341</td>
<td>MAP-Day1 (0.0297)</td>
</tr>
<tr>
<td>F102-Day3</td>
<td>0.0341</td>
<td>MAP-Day1 (0.0293)</td>
</tr>
<tr>
<td>PF-Day0</td>
<td>0.0324</td>
<td>MAP-Day1 (0.0309)</td>
</tr>
<tr>
<td>PRISM12</td>
<td>0.0290</td>
<td>PaO2-Day2 (0.0275)</td>
</tr>
</tbody>
</table>

4.3 Overall Classification Performance

Table 1 shows the prediction performance (area under receiver operating characteristic curve (AUROC) and area under precision-recall curve (AUPRC)) of all methods. The results are averaged over 5 random trials of 5-fold cross validation. We observe that for both tasks, all deep learning models perform better than baseline models. The best performance of deep learning models is achieved by the combination model, which use both DNN and GRU to handle static and temporal input variables, respectively. Our interpretable mimic approach achieves similar (or even slightly better performance) as deep models. We found that Pipeline 1 yields slightly better performance than pipeline 2. For example, Pipeline 1 and 2 obtain AUROC score of 0.7898 and 0.7670 for MOR task, and 0.7889 and 0.7799 for VFD task, respectively. Therefore, we use pipeline 1 model in the discussions in Section 4.4.

4.4 Interpretations

Next, we discuss a series of solutions to interpret Gradient Boosting trees in our mimic models, including feature importance measure, partial dependence plots and important decision rules.

4.4.1 Feature Influence

One of the most common interpretation tools for tree-based algorithms is feature importance (influence of variable). The influence of one variable \( j \) in a single tree \( T \) with \( L \) splits is based on the numbers of times when the variable is selected to split the data samples. Formally, the influence \( I_{f_j}(T) \) is defined as

\[
I_{f_j}(T) = \sum_{l=1}^{L-1} I_l^2 \mathbb{I}(S_l = j),
\]

where \( I_l^2 \) refers to the empirical squared improvement after split \( l \), and \( \mathbb{I} \) is the identity function. The importance score of GBT is defined as the average influence across all trees and normalized across all variables. Although importance score is not about how the feature is actually used in the model, it proves to be a useful metric for feature selection.

Table 2 shows the most useful features for MOR and VFD tasks, respectively, from both GBT and the best GBTmimic models. We find that some important features are shared by several models, e.g., MAP (Mean Airway Pressure) at day 1, \( \Delta PF \) (Change of PaO2/FIO2 Ratio) at day 1, etc. Besides, almost all the top features are temporal features. Among the static features, PRISM (Pediatric Risk of Mortality) score, which is developed and commonly used by doctors and medical experts, is the most useful static variable. As our mimic method outperforms original GBT significantly, it is worthwhile to investigate which features are considered as more important or less important by our method.

Figure 4 shows the individual (i.e. feature importance of a single feature) and cumulative (i.e. aggregated importance of features sorted by importance score) feature importance of the two tasks. From this figure, we observe that there is
Figure 4: Individual (with left y-axis) and cumulative (with right y-axis) feature importance for MOR (top) and VFD (bottom) tasks. x-axis: sorted features.

Figure 5: Feature importance for static features and temporal features on each day for two tasks.

no dominant feature (i.e. feature with high importance score among all features) and the most dominant feature has a importance score less than 0.05, which implies that we need more features for obtaining better predictions. We also noticed that for MOR task, we need less number of features compared to the VFD task based on the cumulative feature importance scores (Number of features when cumulative score > 0.8 is 41 for MOR and 52 for VFD).

We show the aggregated feature importance scores on different days in Figure 5. The trend of feature importance for GBTmimic methods is Day 1 > Day 0 > Day 2 > Day 3, which means early observations are more useful for both MOR and VFD prediction tasks. On the other hand, for GBT methods, the trend is Day 1 > Day 3 > Day 2 > Day 0 for both the tasks. Overall, Day-1 features are more useful across all the tasks and models.

4.4.2 Partial Dependence Plots

Visualizations provide better interpretability of our mimic models. We visualize GBTmimic by plotting the partial dependence of a specific variable or a subset of variables. The partial dependence can be treated as the approximation of the prediction function given only a set of specific variable(s). It is obtained by calculating the prediction value by marginalizing over the values of all other variables.

One-way Partial Dependence Table 2 shows the list of important features selected by our model (GBTmimic) and GBT. It is interesting to study how these features influence the model predictions. Furthermore, we can compare different mimic models by investigating the influence of the same variable in different models. Figure 6 shows one-way partial dependence scores from GBTmimic for the two tasks. The results are easy to interpret and match existing findings. For instance, our mimic model predicts a higher chance of mortality when the patient has value of PH-Day0 below 7.325. This conforms to the existing knowledge that human blood (in healthy people) stays in a very narrow pH range around 7.35 - 7.45. Base blood pH can be low because of metabolic acidosis (more negative values for base excess), or from high carbon dioxide levels (ineffective ventilation). Our findings that pH and Base excess are associated with higher mortality corroborate clinical knowledge. More useful rules from our mimic models can be found via the partial dependence plots, which provide deeper insights into the results of the deep models.

Two-way Partial Dependence In practical applications, it would be more helpful to understand the interactions between most important features. One possible way is to generate 2-dimensional partial dependence for important feature pairs. Figure 7 demonstrates the 2-way dependence scores of the top three features used in our GBTmimic model. From the left figure in Figure 7, we can see that the combination of severe metabolic acidosis (low base excess) and big reduction in PF ratio may indicate that the patients are developing multiple organ failures, which leads to mortality (area in red). However, big drop in PF ratio alone, without metabolic acidosis, is not associated with mortality (light cyan). From the middle figure, we see that low PH value from metabolic acidosis (i.e., with low base excess) may lead to mortality. However, respiratory acidosis itself may not be bad, since if pH is low but not from metabolic, the outcome is milder (green and yellow). The rightmost figure shows that a low pH with falling PF ratio is a bad sign, which probably comes from a worsening disease on day 1. But a low pH without much change in oxygenation is not important in
mortality prediction. These findings are clinically significant and has been corroborated by the doctors.

4.4.3 Top Decision Rules

Another way to evaluate our mimic methods is to compare and interpret the trees obtained from our models. Figure 8 shows two examples of the most important trees (i.e., the tree with the highest coefficient weight in the final prediction function) built by interpretable mimic learning methods for MOR and VFD tasks. Some observations from these trees are as follows: Markers of lung injury such as lung injury score (LIS), oxygenation index (OI), and ventilator markers such as Mean Airway Pressure (MAP) and PIP are the most discriminative features for the mortality task prediction, which has been reported in previous work. However, our selected trees provide more fine-grained decision rules. For example, we can study how the feature values on different admission days can impact the mortality prediction outcome. Similar observations can be made for the VFD task. We notice that the most important tree includes features, such as OI, LIS, Delta-PF, in the top features for VFD task, which again agrees well with earlier findings.

5 Summary

In this paper, we proposed a simple yet effective interpretable mimic learning method to distill knowledge from deep networks via Gradient Boosting Trees to learn interpretable models and strong prediction rules. Our preliminary experimental results show that our proposed approach can achieve state-of-the-art prediction performance on Pediatric ICU dataset, and can identify features/markers important for mortality and ventilator free days prediction tasks. For future work, we will build interactive interpretable models which can be readily used by clinicians.
Acknowledgment

This work is supported in part by NSF Research Grant IIS-1254206 and IIS-1134990, and USC Coulter Translational Research Program. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government.

References


7. Schulam P, Wigley F, Saria S. Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery. 2015;.


Designing a Clinical Data Warehouse Architecture
to Support Quality Improvement Initiatives

John D. Chelico, MD¹, Adam B. Wilcox, PhD², David K. Vawdrey, PhD³,
Gilad J. Kuperman, MD PhD³

¹Northwell Health, Manhasset, NY; ²University of Washington, Seattle, WA;
³NewYork Presbyterian Hospital, New York, NY

Abstract

Clinical data warehouses, initially directed towards clinical research or financial analyses, are evolving
to support quality improvement efforts, and must now address the quality improvement life cycle. In
addition, data that are needed for quality improvement often do not reside in a single database, requiring
easier methods to query data across multiple disparate sources. We created a virtual data warehouse at
NewYork Presbyterian Hospital that allowed us to bring together data from several source systems
throughout the organization. We also created a framework to match the maturity of a data request in the
quality improvement life cycle to proper tools needed for each request. As projects progress in the Define,
Measure, Analyze, Improve, Control stages of quality improvement, there is a proper matching of
resources the data needs at each step. We describe the analysis and design creating a robust model for
applying clinical data warehousing to quality improvement.

Introduction

The advent of electronic health record systems led to the emergence of backend electronic data
repositories of patient and provider information. As repositories grew, clinical data warehouses were
created and optimized for retrospective analysis on patient populations. This analysis has historically been
focused on either clinical research or financial and organizational queries. As more information has been
incorporated in healthcare data warehouses, the potential for data warehouses to address other tasks have
been suggested that focus more on direct patient care improvement.¹² However, it is often unclear how to
best make use of data in warehouses for quality improvement activities. In addition, there are many tools
for accessing and analyzing data in a data warehouse, but their appropriate application can be difficult.
We present a case study of a clinical data warehouse architecture that has been implemented at Columbia
University Medical Center to focus on the appropriate and efficient use of resources for quality
improvement. We modeled our data warehouse and associated tools around a framework that follows the
various phases of quality improvement life cycle. We examine the quality improvement lifecycle used for
improving care at an academic medical center and describe how we leverage the functions of clinical data
warehouse tools according to the maturity of a particular quality improvement project. Finally, we
illustrate the application of this architecture across the breadth of quality improvement initiatives at our
institution.

Background

The Clinical Data Warehouse (CDW) at Columbia University Medical Center (CUMC) of NewYork
Presbyterian Hospital (NYP) was originally created in 1994 by the Department of Biomedical Informatics
in conjunction with the Columbia University Office of Clinical Trials to primarily support clinical
research.³ The data in the warehouse was largely populated by clinical encounter data from the home
grown WebCIS electronic health record (EHR) system backend clinical data repository (CDR).⁴ With this
information from the repository, the warehouse could allow queries of all data available in electronic form
that was typically used in the care of patients. The CDR was created as an event based transactional
database optimized for data retrieval of single patient data, while the CDW was created as an entity based
analytical database optimized for cross patient (population) based retrieval of data. As the use of the CDW grew by clinical researchers, it was soon observed that the data could be used for other purposes such as financial, administrative and, clinical quality improvement in the hospital.

Concurrently there were two important developments at NYP that affected the data warehousing approach. First, NYP organized a quality improvement organization with teams trained in process improvement throughout the organization. Quality improvement was implemented using the Six-Sigma methodology to decrease variation in organizations tasks by carefully examining facts and data in improving existing processes. The application of data to quality improvement required adjustment of the data organization in the warehouse and additional tools to provide data in different workflows related to quality improvement implementation. Second, NYP completed its migration from the legacy WebCIS EHR (which was primarily used for data access to multiple data sources that were integrated in the repository) to a commercial EHR that was implemented across the system. With this new EHR implementation and federal incentive programs came an increased focus on data entry into the EHR. This affected the data warehouse by adding a significant source of new data to the existing design.

As a result of these changes, we sought to better understand how the changes could best be addressed by adaptations to the data warehouse architecture, and to create an effective model for applying data resources to quality improvement efforts.

**Methods: Analysis of Existing Issues**

We performed four analyses to understand perceptions and factors relevant to the current data warehouse and its potential redesign. Two analyses focused on the customers of the warehouse, while two analyses focused on its structure and processes. We used a mixture of an external stakeholder analysis, key user interviews and an analysis of data requests to formulate a guide a new vision for clinical data warehousing at Columbia University Medical Center and the whole NewYork Presbyterian Hospital system.

We first performed a stakeholder analysis among 6 important stakeholders across the institution, who sponsored or facilitated projects that relied on the data that was stored with the clinical data warehouse. These individuals included leaders in the information systems and quality improvement organizations, who had a broad understanding of the data requirements for quality improvement and the need for robust systems to provide that data. They were members of a committee called the Clinical Quality and IT (CQIT) Committee, which was formed with the mission to address key NYP goals for regulatory reporting, patient safety, pay for performance measures, and quality of care delivered through the use of information technology. Along with other efforts to improve the process at the hospital through the use of information technology, they participated in an analysis to inform how the clinical data warehouse could assist in the mission of the quality improvement administrators in the institution. During prioritization via a voting mechanism, a few key themes arose that included being able to track specific safety measures, access to data for analysis, quality reporting and research, the ability to integrate data across applications for reporting purposes, and ability to extract data from clinical notes for quality purposes. The committee felt that these themes could be better addressed through two key subgroups that concentrate on excellence in data warehousing and increased use of structured documentation. The data warehousing subgroup, through a gap analysis of user needs, found that the institution required a more effective method to bring together disparate clinical data sources around NYP and required a more robust way to manage user requests for data.

Next we performed a needs analysis by interviewing key trained quality performance experts with in the institution to better understand their data needs from the warehouse. We concentrated our efforts along six key quality improvement initiatives (Table 1) set forth by NYP senior leadership, and interviewed a leader from each initiative. Our goal was to determine how information technology through data warehousing could help these specific efforts at the institution’s hospitals, and how well the needs were currently being met. Through semi-structured interviews we gathered information from these users in
regards to the maturity of the initiatives, the data needs of the initiative, IT system dependence, and challenges faced in accomplishing goals (see Table 2). The institution already had data used in monitoring these projects, however the data was manually gathered and processed into spreadsheet-based static reports.

Table 1: Quality initiatives pursued at NYP during the analysis period.

<table>
<thead>
<tr>
<th>Initiative</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication Reconciliation</td>
<td>Goal is to create a standardized method to accurately and completely reconcile patient medications between admission and discharge in both the inpatient and outpatient settings.</td>
</tr>
<tr>
<td>Blood Stream Infections</td>
<td>Goal is to decrease the rate of central line associated blood stream infections at NYP and to comply with mandatory New York State reporting of central line associated events.</td>
</tr>
<tr>
<td>Transplant</td>
<td>Goal is to have 100% compliance with the United Network of Organ Donors (UNOS) procedures at NYP and create a robust program for NYP transplant patient quality improvement.</td>
</tr>
<tr>
<td>Patient Verification</td>
<td>Goal is to implement a process to reduce the number of patient identification errors at NYP by ensuring the patients are properly matched to the care they are given.</td>
</tr>
<tr>
<td>NYSSIPP &amp; Pre-Op Antibiotics</td>
<td>Goal is to implement and monitor the New York State Surgical and Invasive Procedure Protocol in order to eliminate wrong patient, wrong site, wrong side and wrong invasive procedure errors at NYP. Goal is also to reduce the rate of post-operative infection at NYP with the proper use and timing pre-operative prophylactic antibiotic administration.</td>
</tr>
<tr>
<td>Pressure Ulcers</td>
<td>Goal is to reduce the prevalence rates of pressure ulcers at NYP by implementing a rigorous staff protocol and acquiring new patient surfaces (beds) to prevent the formation of pressure ulcers.</td>
</tr>
</tbody>
</table>

Table 2: Themes used in semi-structured interviews with quality performance experts.

<table>
<thead>
<tr>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of process issues</td>
</tr>
<tr>
<td>Clarity of vision of over-all approach</td>
</tr>
<tr>
<td>Challenges</td>
</tr>
<tr>
<td>Near term IT Needs</td>
</tr>
<tr>
<td>Strategic Implications</td>
</tr>
<tr>
<td>Systems / Dependencies</td>
</tr>
</tbody>
</table>

These interviews identified a few key issues arose of why the quality improvement needs of the institution were not being met. First, frequently the data needed by quality performance experts was not available within the CDW and that the data needed for these projects was mainly found in vendor based transactional systems. Second, we found that when the data was available they still needed help in defining what their data needs really were. Third, there was no way for the data experts to query across
transactional systems with the IT tools available to them. Finally, we noted that data needs changed according to the maturity of the quality improvement project. The more mature the quality improvement strategy was the more robust IT solution had to be. Still, in each initiative there was a common thread of how data was discovered, validated, and then used to monitor a particular quality improvement marker.

We also performed an analysis of the warehouse itself to identify any structural or technical characteristics that may have been root causes of the concerns raised by the CQIT review. This analysis reviewed requirements and technical factors that were seen by the data warehouse leadership as most significant in their relationship to the CQIT concerns. This analysis identified that the primary challenges were regarding data integration. First, assembling and integrating all administrative, financial, and clinical systems in one system was an immense task. While some of the institution transactional data repositories were well understood and map to a common terminology (the Medical Entities Dictionary, or MED), other newer commercial systems were less understood, and fewer tools were available for navigating the terminology. Prospective mapping large backend vendor based clinical data repositories to a common representation in the clinical data warehouse was a very time consuming goal, beyond the resources of the data warehouse team. Ideally vendors can transfer data from their systems through HL7 messaging and interfaces to the data warehouse, however we found that this was not always well supported by external systems. On analysis of the different data repositories around the institution we found several localized experts centered around different home grown and vendor based systems. This was exacerbated further by having two different patient institutional systems between the university hospitals of Columbia and Weil Cornell. In such an environment, consumers with new data needs had to either “hunt around” for the appropriate person to help them, or would try to push all requests through whoever in the past had successfully met the data need.

The findings of the CQIT committee and the focused quality performance expert interviews further justified the need for a clinical data warehouse that served the needs of the quality improvement initiatives at the institution. Moreover, we felt it was also important to maintain and perhaps improve the capabilities that were already being supported in clinical research and administration. Therefore, we performed a process analysis of the current activities of the CDW team by reviewing requests submitted to the CDW, and collecting feedback from the CDW analysts. This review included data requests made formally through a data request submission process, and informally through requesters seeking instruction directly through CDW leaders. We identified the maturity of each data request in terms of the existing access to the data, the roles of the users of the data, and how the data were integrated with the users’ workflow. We also examined the sources of data used for the data requests, and the overall data flow from our source systems to their eventual use in our CDW. We then grouped different data uses and data sources into common types, and qualitatively identified and analyzed themes among the sources and uses.

The CDW user requests have classically focused on the clinical researchers at CPMC. More recently the analysts observed an increase in requests from administrative and quality improvement personnel at the institution. Requests differ from the classic research oriented queries in that they require more time commitment and expertise from the CDW staff. Requests ranged from one-time queries to quarterly and even daily reporting to administrative systems. They also required bringing together financial, clinical, and administrative data into one query. The CDW team was spending significant time running ad-hoc queries in a one time or recurrent fashion based on the user. Additionally almost all of the ad-hoc inquiries where done without looking to building upon previously understood data requests. The CDW looked for an infrastructure for managing repeated common requests and users that keep requesting more complex data mining efforts.

**Results: System Description and Design**

From these four analyses, we were able to identify three themes that influenced the redesign of our clinical data warehouse. These themes were 1) structural changes in our data infrastructure created different needs for data integration; 2) ad hoc queries were a foundational method for accessing data and
Data Integration

A dominant theme that emerged from the stakeholder and structural analyses was regarding the need for improved data integration to support customer needs. Because we were in an academic medical center, the data warehouse had to consider the goals and data needs of both the medical school and hospital – the medical school needs were focused more on education and research, while the hospital focus was more on patient care, clinical quality, and financial measures. The designed warehouse architecture also needed to be responsive to user requests with minimal resources. The key to this architecture was handling user requests and maintaining support for research and quality improvement projects.

As mentioned above, the main data source for the CDW was the institution’s clinical data repository, which had as its sources data from multiple ancillary systems that were interfaced to a central database, and the more recent institutional electronic health record that included computerized physician order entry and a clinical documentation system. Because the EHR was intended to replace many functions of the legacy WebCIS application and CDR, the data from the EHR were not modeled and stored in the repository, and many data sources that fed to the repository also were stored directly in the EHR (e.g., lab data). This change in the primary data source for the EDW was important for the new design, as the warehouse needed to accommodate data integration that was previously done in the repository and using the MED. Because we did not have the resources for full data integration between the systems, and the EHR data represented a new data type, we focused on the extract process to simplify access to the data and modeled or transformed the data based on need (e.g., late-binding).

For data extraction, we were able to use standardized database management tools to make the data accessible in the CDW without actually consolidating the data sources by creating a virtual clinical data warehouse. We first replicated the source application’s backend database management system in real time, using standard database replication tools. This minimized the effect on the front-end application performance by not burdening the application to feed data to our warehouse directly. We then used existing database management system integration services to link the replicated database with existing tables in the CDW. These integration services supported both ad hoc analytic queries and extracting data from EHR tables to populate data marts in the CDW. Data marts contain data that is pertinent to its area of interest and hence only contains a focused subset of the source replicated databases, and required data transformation and modeling. Based on the demand of the data needs, we could focus efforts to asynchronous ad hoc queries or have more recurring and even synchronous reporting and analysis through content optimized data marts. Different data marts could be updated at defined intervals or in real time to support online analytical processing. Database integration services for both ad hoc queries and updates to centralized data marts allowed more rapid availability of the data for requesters. We found that this approach was most effective for supporting ad hoc queries, and could be effective in building data marts to support periodic data reports.

Data Access and Ad Hoc Queries

Themes regarding access and ad hoc queries were identified primarily from the process and needs analyses. We noted that requests from our users, whether they were researchers, clinical providers, or operational analysts, were all initially performed as ad hoc queries. Even when the data were needed in a complicated report or dashboard, the data were first extracted and verified using provisional queries, to check the data definitions, availability, and quality. Research queries were generally all ad-hoc queries, with few examples of where the queries were run repeatedly at intervals. While we did provide other functions within the CDW such as generated reports and data feeds to separate applications, these
requests were enhancement requests for more robust delivery of data that was originally requested as an ad hoc preliminary query. Understanding that the data requests all started as ad hoc queries allowed us to identify a pattern of how data requests evolved, and allowed us to design around optimizing data access for ad hoc queries separately from modeling for reporting. When a query is requested on a recurrent basis, often to identify trends in indicators, data marts could then be built and extract-transform-load processes were implemented to aid in the data modeling and consolidation process (see Figure 1). This was both to increase the efficiency of the end-user data extract, and also because the data modeling was better defined for these more mature requests. Ongoing requests made by applications or analytical systems, such as our business intelligence system, were handled in a similar fashion. Data marts were created in the warehouse and either used directly by the front end systems or extracted to the system database.

![Virtual Clinical Data Warehouse Design](image)

**Figure 1. Virtual data warehouse design.**

We also found a multitude of data visualization tools and services that required different levels of information accessibility from our clinical data warehouse. After careful examination of our requests we found that they could be divided into four distinct categories based upon the frequency or urgency of data required. These categories were as follows:

1) Ad-hoc Direct SQL Queries
2) Recurring Static Reporting
3) Online Analytical Processing
4) Point of Care Registry Functions
Each of these categories provided for a predefined period of time for a new inquiry to our data warehouse tables. From an ad-hoc query that maybe done once to more timely queries required for point of care reporting, each of these categories would require data updated at predefined intervals driven by the frequency of the reporting task.

Quality Improvement Cycle

The needs and process analyses identified the changing user needs that were related to the quality improvement life cycle. The fourth lesson from the needs analysis was that data needs changed according to the project maturity in the life cycle, and the process analysis identified multiple methods for supplying data. We were able to map our categorization of data warehouse data methods to the quality improvement cycle initiatives followed at our institution. The approach followed the DMAIC framework: DEFINE, MEASURE, ANALYZE, IMPROVE and, CONTROL, each representing different stages of the quality improvement life-cycle, each with differing data needs. At the DEFINE stage the user asks what type of data is needed and if it exists. With the MEASURE stage static reporting can monitor trends in data over time in order to find a baseline level for future measurements. During the ANALYZE stage data is further looked at from different angles and viewpoints in order to find the best place for intervention to make the biggest impact. Interventions are put into place and the data is used to help care givers in making the right decisions at the point of care during the IMPROVE stage. Finally, at the CONTROL stage, process improvement is incorporated in the workflow of an organization and data driven decision making is becomes part of the standard of care.

The mapping of the data access methods to the quality improvement cycle was as follows. New projects in the DEFINE phase always started with a direct database query to establish the availability and validity of the data. As projects matured, recurring automated queries would help further define the project and identify trends over time in the MEASURE phase. Deeper analysis of the data would use more advanced online analytical processing tools to find factors that correlated with the quality improvement task at hand in the ANALYZE phase. Data-driven dashboards then can bring information from real-time data marts to clinicians at the point of care so they could monitor key variables in the IMPROVE phase. As the use of the key variables was better understood, it could be integrated into workflow of the clinicians with real time decision support in the CONTROL phase (see Table 3).

<table>
<thead>
<tr>
<th>Quality Improvement Stage</th>
<th>Data Need</th>
<th>Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINE</td>
<td>Research</td>
<td>Ad hoc queries</td>
</tr>
<tr>
<td>MEASURE</td>
<td>Management reports</td>
<td>Automated queries / reports</td>
</tr>
<tr>
<td>ANALYZE</td>
<td>Operational reports</td>
<td>On-line analytical processing</td>
</tr>
<tr>
<td>IMPROVE</td>
<td>Point-of-care reporting</td>
<td>Dashboards</td>
</tr>
<tr>
<td>CONTROL</td>
<td>Decision support</td>
<td>Alerts / automated orders</td>
</tr>
</tbody>
</table>

While this mapping to the DMAIC cycle was critical in matching data needs to appropriate tools for quality improvement initiatives. Because of the data-intensive nature of this process, leaders of quality improvement initiatives would often approach the clinical data warehouse managers to meet their data needs. This led to consideration of how to optimize the data warehouse around quality, rather than its traditional role as a retrospective clinical research tool. The phases were not always concretely defined in the data requests; however, the task being performed by the quality expert at the time of the data warehouse inquiry typically defined the complexity of the data warehouse storage and the tools used to...
access it. It also defined our expectations of how the data needs would evolve as the quality improvement initiative matured. As a project evolved from a “define” stage to “improve” or “control” stage, requirements for data timeliness and access would increase accordingly. This further classification of user needs based upon the maturity of the quality improvement initiative created a framework for prioritization of user requests and proper data mart utilization. Using this model, the CDW team sought to create a clinical data warehouse architecture that focused on the types of data required for each stage of quality improvement in order to better serve the quality improvement efforts of NYP.

Results –Diabetes Mellitus Example

We illustrate the application of this clinical data warehouse architecture supporting the stages of the quality improvement life cycle using a quality improvement initiative concentrating on the outpatient clinic population of patients with diabetes mellitus at the Columbia University Medical Center.

Ad-hoc SQL Queries (Define): For several years healthcare providers from the NewYork-Presbyterian Ambulatory Care Network intermittently requested information from our clinical data warehouse on their patients with diabetes mellitus. Each request was done on a case by case basis and customized for a particular purpose in mind, whether it was for clinic research or administrative reporting.

Static Reporting (Measure): Starting in 2006, some of these reports became more formalized and a process for automated querying and delivery of data was setup by the clinical data warehouse staff. With the help our data warehouse staff, medical directors would monitor the patients with diabetes in their clinic population and provide reports to their physicians regarding timeliness and control of laboratory values including Hemoglobin A1C, LDL, and Microalbumin. On a quarterly basis data would be compiled and delivered via “flat file” text files to the healthcare analysts in the Ambulatory Care Network for processing and reporting. This data would ultimately be delivery to healthcare providers in a paper based document. Aggregate data would be provided on both the individual physicians and clinic locations on how goals were being met for each laboratory monitor. Physicians would also receive a registry of the patients included in their calculations with associated demographic and laboratory values.

Online Analytical Processing (Analyze): As these reports became more and more utilized, the need for more frequent reporting was evident. Much of the data manipulation needed to make such reports required a great deal of manual effort and the data would sometimes be delivered when it was already temporally unusable. Additionally as medical directors sought to use this data to manage quality improvement interventions they found it very burdensome to monitor their efforts in this way.

With the goal of having timelier reporting to healthcare providers we migrated the process of data gathering into data-marts and automated the data analysis through online analytical processing tools. From various disparate data sources we built a data-mart of diabetic patient demographics, outpatient visit history, laboratory, medication, and provider information. We then created a dynamic management reporting tool using our business intelligence system to provide timely patient and provider tracking tools used at the point of care. With this new method we were able to scale up to our users needs of more frequent reporting on clinic patients with diabetes mellitus.

Dashboards-Registries (Improve): Future improvements to clinical care in diabetes will turn to our clinical data warehouse to provide instant access to patient and provider information regarding diabetic patient care. With our architecture in place we can update data-mart content with a frequency required by user needs. As new tools such as healthcare provider dashboards and chronic disease registries are implemented to manage the population of patients with diabetes in our institution, the data warehouse can provide the platform for future expansion.

Applications (Improve): As interventions are defined to improve the care of patients with diabetes they must be integrated into the transactional electronic health record applications used in the institution. New decision support alerts and reminders can keep the patient care providers up to date on what the needs are of a particular patient.
We observed in our outpatient diabetes mellitus quality improvement initiative the maturity of the project dictated the resources needed from the clinical data warehouse and its users. By planning the proper approach and building upon prior less sophisticated reporting methods we were able to create a data-mart architecture and robust data visualization tool that exactly met the needs of our users.

**Discussion**

Using multiple analyses, we were able to identify important themes regarding our data warehouse that led to a specific redesign to support evolving needs. We used virtualization of data sources and data marts to facilitate late-binding of data models and to support ad hoc queries that were foundational to using the warehouse data. Organizing our clinical data warehouse architecture around the quality lifecycle has been an effective approach to meeting the evolving demands of a data warehouse. The fundamental architecture of our virtual clinical data warehouse and associated prioritization strategy around our data visualization tool has brought to our attention a few key findings.

First, our data warehouse prioritization strategy made effective use of our technical resources for a given project. It was imperative that the proper resources were also given to the project from our users perspective. As the maturity of the quality improvement initiative progressed, more resources were needed on the user side to ensure success of the project. While ad-hoc queries required only minimal user interaction, mature projects required greater time commitment and technical proficiency from our users. Additionally, as a project matured, higher level management skills were required of the administrative staff monitoring its progress. As our diabetes project moved from static reporting to online analytical processing, the need for project management skills was evident to coordinate the different users and technical staff for the project.

Second, as we built customized data-marts we needed to keep in mind that it was important to keep data-mart models separated from the data visualization tools. Different tools required different backend data sources. This abstraction of backend supply side and frontend demand side environments of a clinical data warehouse has been described in the University of Michigan Health System. As a project matures different tools will need to interact with the same data-mart. Keeping data-marts as generic as possible facilitates future requirements and expansions to the data model.

Third, defining projects based on maturity helped to ensure that users were provided with appropriate tools for the task. Once we identified the expected progression of a project through the quality improvement life cycle, we were able to avoid situations where we would try to extend tools of one level to different functions as the project evolved. For example, prior to the mapping against the quality improvement model, projects at a Measure stage, using reporting software, might be steered to extend the tools to different functions. We experienced discussions of how to implement reporting tools at the point of care, which would be an expensive and difficult implementation. With the understanding of project evolution, we recognized that projects were expected to evolve, and that we should migrate to the appropriate tools rather than extend the tools at one level.

Some limitations exist in the architecture of a virtual data warehouse. By relying on discretely replicating backend databases of source systems we create a delay in delivery of data to the warehouse. While our data warehouse is not formally used for transactional processes, it is feasible in our model that data be available in a synchronous manner for dashboard style point of care reporting. To solve this we would need to look more toward HL7 messaging and interfaces that would feed data directly from source systems as it is entered. We have recently begun testing of loading data directly into the CDW using HL7 messaging, which would improve the data currency.

Additionally, as with any data warehousing approach, one must be sensitive to the quality of data that is in the warehouse tables. Data can only be as good as what is available in the source systems. At every level of warehouse function, we find that data must be cleaned in order to create meaningful reports to
users. In the future we expect to create a scoring system for data points stored in our warehouse. This way one can not only provide data, but also can impart a level of integrity for data points provided.

While the analysis-derived themes and design were valuable as they were applied to the data warehouse at NewYork Presbyterian Hospital, perhaps their most significant characteristic has been their robustness beyond the institution where they were developed. Since this initial analysis, discovery and application, some of our team members have migrated to different institutions, and we have applied these findings to warehousing and analytics strategies in these other settings. In each case, these important lessons have been effective in creating local strategies for addressing the evolving user needs of clinical data, especially for quality improvement initiatives.

**Conclusion**

We found that by concentrating on the needs of our users we could establish a robust clinical data warehouse that supports the needs of quality reporting initiatives at our institution. By building a clinical data warehouse architecture to support the dynamic needs of data visualization tools available we could scale up resources as user needs required. By creating a framework for user needs based upon the maturity of quality improvement initiatives we could direct users to the proper level of data warehouse tools.

**References**

Integrative Spatial Data Analytics for Public Health Studies of New York State

Xin Chen, MS, Fusheng Wang, PhD
Stony Brook University, Stony Brook, NY

Abstract

Increased accessibility of health data made available by the government provides unique opportunity for spatial analytics with much higher resolution to discover patterns of diseases, and their correlation with spatial impact indicators. This paper demonstrated our vision of integrative spatial analytics for public health by linking the New York Cancer Mapping Dataset with datasets containing potential spatial impact indicators. We performed spatial based discovery of disease patterns and variations across New York State, and identify potential correlations between diseases and demographic, socio-economic and environmental indicators. Our methods were validated by three correlation studies: the correlation between stomach cancer and Asian race, the correlation between breast cancer and high education population, and the correlation between lung cancer and air toxics. Our work will allow public health researchers, government officials or other practitioners to adequately identify, analyze, and monitor health problems at the community or neighborhood level for New York State.

Introduction

Open data initiatives supported by the governments are providing unprecedented information about our health. New York State Cancer Mapping dataset, for example, consists the number of people diagnosed with cancer (cancer counts, 2005-2009) in small geographic areas. New York State data from Statewide Planning and Research Cooperative System (SPARCS) collects patient level detail on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient visit. Such data also provides street level location information for each patient and healthcare facility site. The improved availability of health data combined with improved geospatial analysis and spatial statistics techniques has significant potential to uncover the spatial patterns of diseases in a population and provide insight as to their causes and controls.

Integrative spatial data analytics for public health has a strong focus on locating patients and the agents of disease, studying the community and region level patterns and variations, and assessing demographic, socio-economic, and environmental factors on diseases and human health. In the past, due to limited accessibility of health outcome data, public health studies were often limited at macro scale levels such as county level or ZIP code tabulation areas (ZCTAs), and may not allow public health researchers and health officials to adequately identify most at-risk populations, analyze, and monitor health events at the community or neighborhood level.

One critical challenge for spatial epidemiology and public health research is the isolated datasets with different spatial resolutions, data formats, or data quality. The patient addresses from hospitals, for example, have to be converted into geolocations (latitude, longitude) and then get approximated into a standard geographic identifier (such as census tract or census block group IDs) to protect the privacy of human subjects. Another common issue is the problem of spatial interpolation that combines both point and areal data through the use of area-to-area, area-to-point, and point-to-point covariances. To align spatial data with different geographical granularities, we also need to either aggregate values from small areas into larger ones or vice versa.

Our goal is to integrate fine-grained open health data with a comprehensive set of spatial “exposure” data, which is ranging from levels of various environmental pollutants to the socioeconomic status of persons at risk. We focus on the spatial public health research at the community level and consolidate a variety of spatial datasets into a data warehouse system, supported by a scalable computing infrastructure for spatial data integration and spatial analytics. We take advantage of Hadoop-GIS, a MapReduce based spatial data warehouse system to perform spatial query based data integration. Integrative spatial data analytics is built on top of the integrative spatial data warehouse to support various spatial query types and analytic methods.

In this paper, we verified our methodology for integrative spatial data analytics by linking the New York Cancer Mapping Dataset (Table 1) with three categories of spatial impact indicators (Table 2) generated from representative spatial exposure data sources such as census statistics, geospatial and map data from TIGER (Topologically Integrated Geographic Encoding and Referencing), the American Community Survey (ACS) data, and air toxics data from Environmental Protection Agency (EPA). We first studied spatial distributions and clustering of the cancer risk and
spatial impact factors at the level of census block group or census tract. We then correlated the cancer risk ratios with spatially varying demographic, socio-economic, and environmental factors (Tables 3-5), using both non-spatial correlation analysis (ordinary least square regression) and spatial correlation analysis (geographically weighted regression). The results were consistent with three well-established relationships: that between stomach cancer and Asian and Hispanic race, that between breast cancer and high income and high education population, and that between lung cancer and air toxics. At last, we undertook three case studies to identify the detailed spatial trends for each of these three pairs of relationship separately (Figures 1-3).

Methods

Data Sources

Spatial Public Health Data: Cancer Incidence Data in New York State. There has long been a demand for cancer incidence data at a fine geographic resolution for use in etiologic hypothesis generation, methodological evaluation and teaching. We demonstrated our vision of integrative spatial analytics by linking the New York Cancer Mapping Dataset with a comprehensive set of spatial impact indicators from representative spatial exposure data sources such as census and TIGER data, the ACS data, and air toxics data from EPA.

The cancer data set consists of observed counts for 23 anatomic sites of cancer at the neighborhood scale, diagnosed between 2005 and 2009 (Table 1). The data include 524,503 diagnoses of cancer distributed across 13,823 Census block groups with an average population of about 1,400. A census block group is an area containing about 1,000 to 2,000 people as defined by the US Census. Cancer data are reported for a five-year time period because the number of cases in single years can vary dramatically, particularly for outside metropolitan areas.

Besides the observed counts of diagnosed cancers, expected counts are also calculated using the indirect standardization method, adjusted for sex and 5-year age groups up to 85+, using the 2010 census counts for New York State. The data set thus contains both observed counts and expected counts per census block group for each of the 23 cancer sites.

Spatial Exposure Data. We linked three categories of spatial impact indicators to the cancer incidence data using both non-spatial correlation analysis and spatial correlation analysis as shown in Table 2. 1) For demographic indicators, we examined the population proportion for different groups of individuals based on self-identified race and ethnicity, including Asian, Hispanic, White, Black, Pacific Islander, and American Indian. 2) For socio-economic indicators, we examined the number of persons per household, the proportion of poverty population and the population with less than high school education. 3) For environmental indicators, we examined 10 most frequently found air toxics with national average cancer risk greater than one in a million.

The demographic indicators here were generated from the 2010 United States Census and the socio-economic indicators were from the 2006-2010 American Community Survey. The environmental indicators for New York State were generated from Environmental Protection Agency (EPA) National-scale Air Toxins Assessment. As the EPA air toxics data were at census tract level, we aggregated lung cancer counts per census block group into the counts per census tract.

Analyses

Relative Risk (RR). We calculated RR through dividing observed counts by expected counts per census block group for different cancer categories (Table 1). The RR estimates provide useful information about how common cancer incidence in a specific location is as compared to the global baseline.

Spatial Clustering. To test whether there is spatial autocorrelation of the cancer RR, we used Moran’s I (Tables 1 and 2). Moran’s I is a widely used global cluster test, which determines the degree of clustering or dispersion within a data set. The resulting values may range from 1 (perfect correlation), 0 (complete spatial randomness) to -1 (perfect dispersed). For the cancer incidence data, a positive spatial autocorrelation means that the areas with high cancer RR are close to other areas with high cancer RR.

To assess colocation (= spatial correlation) between cancer incidence and each of the spatial impact indicators, we calculated the bivariate Moran’s I for each spatial impact indicator with the cancer RR (Tables 3 and 4). Such bivariate measure of spatial autocorrelation relates the value of a cancer RR at a location to that of a spatial indicator at neighboring locations, as a straightforward generalization of the concept of spatial autocorrelation. This measure of spatial autocorrelation shows the association between cancer incidence at a given location and the local indicator value in neighboring area. The Moran’s I index and the bivariate Moran’s I index were conducted in GeoDa (v 1.6.7).
In addition to observed and expected counts, the New York Cancer Mapping Dataset also includes an indicator variable used to highlight block groups with unusually high or low cancer incidence, as determined using the spatial scan statistic. The spatial scan statistic is a local cluster test, which detects local clusters with statistically significant elevated or deficit risk of diseases. A block group is defined as a high incidence area if 1) it was included in the most likely high incident rate cluster detected by the spatial scan statistic, or 2) it was included in a non-overlapping secondary cluster, and 3) the observed rate was at least 50% higher than the expected rate. A block group was defined as a low incident area in the same manner.

**Spatial Regression.** To identify potential correlations between diseases and spatial impact indicators, we assessed both non-spatial and spatial correlation. Ordinary least square regression analyses (OLS) was used to determine non-spatial correlation between cancer RR and each of the spatial impact indicators. We then used Geographically Weighted Regression (GWR) to assess the spatial trends of relevant indicators at a local level.

OLS is a linear regression method that closely fits a function by minimizing the sum of squared errors. To determine potential candidate indicators, we evaluated several possible indicator combinations that form a properly specified OLS regression model. GWR is a local form of linear regression used to model spatially varying relationships. To evaluate the correlation between cancer incidence and spatial impact indicators accounting for data in surrounding areas, we conducted GWR with a selected indicator combination based on OLS results. Specifically, a fixed kernel type function was used to calculate the GWR regression coefficients. The extent of the kernel is determined using the Akaike Information Criterion (AICc). The OLS, GWR and resulting choropleth maps (Figures 1-3) were conducted in ArcGIS (v 10.3).

**Table 1.** Statistics of New York Cancer Mapping Dataset at Census Block Group Level, 2005-2009.

<table>
<thead>
<tr>
<th>Category of Cancer</th>
<th>Sum of Cases</th>
<th>Relative Risk (RR)</th>
<th>Moran's I Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>Total</td>
<td>524,503</td>
<td>0.18</td>
<td>17.14</td>
</tr>
<tr>
<td>Prostate</td>
<td>78,162</td>
<td>0</td>
<td>12.08</td>
</tr>
<tr>
<td>Female breast</td>
<td>72,296</td>
<td>0</td>
<td>10.89</td>
</tr>
<tr>
<td>Lung and bronchus</td>
<td>67,217</td>
<td>0</td>
<td>21.83</td>
</tr>
<tr>
<td>Colon and rectum</td>
<td>49,801</td>
<td>0</td>
<td>21.50</td>
</tr>
<tr>
<td>Bladder, including in situ</td>
<td>25,134</td>
<td>0</td>
<td>13.68</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>22,279</td>
<td>0</td>
<td>8.93</td>
</tr>
<tr>
<td>Uterus</td>
<td>17,194</td>
<td>0</td>
<td>79.63</td>
</tr>
<tr>
<td>Kidney and renal pelvis</td>
<td>16,371</td>
<td>0</td>
<td>39.27</td>
</tr>
<tr>
<td>Thyroid</td>
<td>15,109</td>
<td>0</td>
<td>12.75</td>
</tr>
<tr>
<td>Leukemia</td>
<td>14,091</td>
<td>0</td>
<td>25.05</td>
</tr>
<tr>
<td>Pancreas</td>
<td>13,927</td>
<td>0</td>
<td>54.31</td>
</tr>
<tr>
<td>Oral cavity and pharynx</td>
<td>10,799</td>
<td>0</td>
<td>16.50</td>
</tr>
<tr>
<td>Stomach</td>
<td>9,285</td>
<td>0</td>
<td>36.70</td>
</tr>
<tr>
<td>Liver and intrahepatic bile duct</td>
<td>8,342</td>
<td>0</td>
<td>15.63</td>
</tr>
<tr>
<td>Ovary</td>
<td>7,582</td>
<td>0</td>
<td>55.14</td>
</tr>
<tr>
<td>Brain and other nervous system</td>
<td>6,714</td>
<td>0</td>
<td>22.33</td>
</tr>
<tr>
<td>Esophagus</td>
<td>5,467</td>
<td>0</td>
<td>17.13</td>
</tr>
<tr>
<td>Larynx</td>
<td>4,179</td>
<td>0</td>
<td>23.65</td>
</tr>
<tr>
<td>Soft tissue</td>
<td>3,385</td>
<td>0</td>
<td>23.83</td>
</tr>
<tr>
<td>Testis</td>
<td>2,690</td>
<td>0</td>
<td>33.61</td>
</tr>
<tr>
<td>Bone and joint</td>
<td>1,026</td>
<td>0</td>
<td>61.12</td>
</tr>
<tr>
<td>Mesothelioma</td>
<td>979</td>
<td>0</td>
<td>60.42</td>
</tr>
<tr>
<td>Nasal cavity and nasopharynx</td>
<td>689</td>
<td>0</td>
<td>84.97</td>
</tr>
</tbody>
</table>

* Significant at 1% confidence interval
Results

In this section, we verified our methodology for integrative spatial analytics by linking cancer incidence data (Table 1) with three categories of spatial impact indicators from spatial exposure data (Table 2). The results were consistent with three well-established relationships: the correlation between stomach cancer and Asian race, the correlation between breast cancer and high education population, and the correlation between lung cancer and air toxics. Our results also included detailed spatial trends between cancer RR and spatial impact indicators through three case studies.

Table 2. Statistics of Demographic, Socio-economic, and Environmental Indicators from Spatial Exposure Data.

<table>
<thead>
<tr>
<th>Spatial Impact Indicators</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean (St. Dev.)</th>
<th>Moran’s I Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Indicators</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian*</td>
<td>0</td>
<td>0.96</td>
<td>0.07 (0.12)</td>
<td>0.86*</td>
</tr>
<tr>
<td>White*</td>
<td>0</td>
<td>1</td>
<td>0.68 (0.31)</td>
<td>0.91*</td>
</tr>
<tr>
<td>Black*</td>
<td>0</td>
<td>0.98</td>
<td>0.15 (0.24)</td>
<td>0.91*</td>
</tr>
<tr>
<td>American Indian*</td>
<td>0</td>
<td>0.95</td>
<td>0.01 (0.02)</td>
<td>0.34*</td>
</tr>
<tr>
<td>Pacific Islander*</td>
<td>0</td>
<td>0.05</td>
<td>0 (0)</td>
<td>0.13*</td>
</tr>
<tr>
<td>Hispanic*</td>
<td>0</td>
<td>0.94</td>
<td>0.16 (0.19)</td>
<td>0.88*</td>
</tr>
<tr>
<td>Socio-economic Indicators</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poverty Population*</td>
<td>0</td>
<td>0.90</td>
<td>0.13 (0.14)</td>
<td>0.54*</td>
</tr>
<tr>
<td>Low Education Population*</td>
<td>0</td>
<td>0.78</td>
<td>0.15 (0.13)</td>
<td>0.60*</td>
</tr>
<tr>
<td>Num. of Persons per Household</td>
<td>0</td>
<td>5.75</td>
<td>2.63 (0.52)</td>
<td>0.70*</td>
</tr>
<tr>
<td>Environmental Indicators**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formaldehyde</td>
<td>6×10⁻⁶</td>
<td>62×10⁻⁶</td>
<td>25×10⁻⁶ (11×10⁻⁶)</td>
<td>0.98*</td>
</tr>
<tr>
<td>Carbon tetrachloride</td>
<td>2×10⁻⁶</td>
<td>4×10⁻⁶</td>
<td>3×10⁻⁶ (0)</td>
<td>0.05*</td>
</tr>
<tr>
<td>PAHPOM</td>
<td>0</td>
<td>26×10⁻⁶</td>
<td>2×10⁻⁶ (2×10⁻⁶)</td>
<td>0.71*</td>
</tr>
<tr>
<td>Chromium VI</td>
<td>0</td>
<td>43×10⁻⁶</td>
<td>2×10⁻⁶ (2×10⁻⁶)</td>
<td>0.84*</td>
</tr>
<tr>
<td>Acetaldehyde</td>
<td>1×10⁻⁶</td>
<td>8×10⁻⁶</td>
<td>4×10⁻⁶ (1×10⁻⁶)</td>
<td>0.94*</td>
</tr>
<tr>
<td>Benzene</td>
<td>1×10⁻⁶</td>
<td>57×10⁻⁶</td>
<td>15×10⁻⁶ (8×10⁻⁶)</td>
<td>0.93*</td>
</tr>
<tr>
<td>Tetrachloroethylene</td>
<td>0</td>
<td>19×10⁻⁶</td>
<td>4×10⁻⁶ (3×10⁻⁶)</td>
<td>0.93*</td>
</tr>
<tr>
<td>Naphthalene</td>
<td>0</td>
<td>33×10⁻⁶</td>
<td>7×10⁻⁶ (4×10⁻⁶)</td>
<td>0.92</td>
</tr>
<tr>
<td>1,3-butadiene</td>
<td>0</td>
<td>14×10⁻⁶</td>
<td>4×10⁻⁶ (2×10⁻⁶)</td>
<td>0.92*</td>
</tr>
<tr>
<td>Arsenic</td>
<td>0</td>
<td>9×10⁻⁶</td>
<td>2×10⁻⁶ (1×10⁻⁶)</td>
<td>0.94*</td>
</tr>
</tbody>
</table>

*Population Proportion among Total Population

** The air toxics data from EPA are at the level of census tract.

Linking Spatial Impact indicators to Health Events

We first modeled all demographic and socio-economic indicators into OLS for both stomach cancer and breast cancer, and then excluded White population for stomach cancer and Black population for breast cancer due to the redundancy issue among different indicators. For the OLS results of stomach cancer RR, we included all indicators except the White population proportion. For the OLS results of breast cancer RR, we included all indicators except the Black population proportion.

Based on OLS results, we then tried different indicator combinations for GWR model and further excluded several indicators due to the global or local multicollinearity issue. As shown in Table 3, we separately chose 5 indicators for GWR model with stomach cancer RR and 3 indicators for GWR model with breast cancer RR. For lung cancer RR (Table 4), 6 out of the 10 air toxics were included in the final OLS model and the 3 air toxics related to lung cancer were included in the final GWR model in the same manner.

The proportion of Asian population was most strongly correlated with stomach cancer RR (Table 3). Among the other significant indicators, the proportions of Black, Hispanic, poverty, and low education population all positively...
influenced stomach cancer RR. Such finding is consistent with the high stomach cancer rates found among Asian-Americans\(^3\).

### Table 3. Summary of Correlation Analysis for Demographic and Socio-economic Indicators.

<table>
<thead>
<tr>
<th>Spatial Impact Indicators</th>
<th>Stomach Cancer RR</th>
<th>Breast Cancer RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Indicators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>1.87* (0.11)</td>
<td>-0.08* (0.04)</td>
</tr>
<tr>
<td>White</td>
<td>-</td>
<td>0.11* (0.02)</td>
</tr>
<tr>
<td>Black</td>
<td>0.82* (0.06)</td>
<td>0.12*</td>
</tr>
<tr>
<td>American Indian</td>
<td>-0.46</td>
<td>-0.67* (0.24)</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>1.70</td>
<td>-10.95* (3.66)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.32* (0.09)</td>
<td>-0.08* (0.03)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Socio-economic Indicators</th>
<th>Stomach Cancer RR</th>
<th>Breast Cancer RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty Population</td>
<td>0.33* (0.12)</td>
<td>-1.26 – 0.31</td>
</tr>
<tr>
<td>Low Education Population</td>
<td>0.57* (0.15)</td>
<td>-1.15 – 0.12</td>
</tr>
<tr>
<td>Num. of Persons per Household</td>
<td>0.01</td>
<td>-0.04* (0.01)</td>
</tr>
</tbody>
</table>

* Significant at 1% confidence interval

Among socio-economic indicators, the proportion of low education population most strongly correlated with breast cancer RR (Table 3). The proportions of different ethnicity groups all negatively influenced breast cancer RR, except the White population proportion. Such findings confirmed the risk factors of higher socio-economic status (SES), race and ethnicity for female breast cancer\(^9\). High SES, which is most often defined by high income and/or high education level, has been linked to an increased risk of breast cancer. This increased risk is not due to the higher SES itself, but rather to differences in risk factors found in women of different education and income levels. For example, compared to women of lower SES, women of higher SES are more likely to 1) have their first child at a later age, 2) have fewer children, 3) use menopausal hormone therapy, 4) use birth control pills, and 5) drink alcohol\(^9\). Since many of these behaviors tend to occur in combination as part of broader lifestyle patterns, the high education can be considered as a direct risk factor for breast cancer.

Most indicators that were significantly correlated with stomach and breast cancer (based on OLS coefficients) also displayed significant colocation relationships (based on bivariate Moran’s I index). However, the bivariate Moran’s I index (spatial correlation) was generally lower than OLS coefficient value (non-spatial correlation), indicating that such correlation was only partly determined by location.

For lung cancer, we examined the 10 most frequently found air toxics with national average cancer risk greater than one in a million (Table 4). The cancer types associated with the air toxics were also listed. We included several air toxics that affected nose cancer, leukemia, and adrenal tumors other than lung cancer for comparison purpose. Among the significant OLS coefficient results, all 3 air toxics related to lung cancer (arsenic, chromium VI, and PAHPOM) were positively correlated with lung cancer. For the air toxics not related to lung cancer, their relationship to lung cancer RR displayed certain randomness. For example, while the carbon tetrachloride (related to adrenal tumors) had the largest positive coefficient value, tetrachloroethylene (related to liver cancer) had a negative coefficient value.

The colocation relationship (Bivariate Moran’s I Index) also displayed inconsistent results as compared to OLS model. For example, while arsenic and chromium VI were positively correlated to lung cancer according to the significantly positive coefficient value, their bivariate Moran’s I indexes were negative which indicated negative spatial correlation.
Table 4. Summary of Correlation Analysis for Environmental Indicators**.

<table>
<thead>
<tr>
<th>Spatial Impact Indicators</th>
<th>Cancer Type</th>
<th>Air Toxics**</th>
<th>Lung Cancer RR</th>
<th>OLS (Std. Error)</th>
<th>GWR</th>
<th>Bivariate Moran’s I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acetaldehyde</td>
<td>Lung Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arsenic</td>
<td>Lung Cancer</td>
<td>0.50×10* (0.14×10⁶)</td>
<td>-0.38×10⁶ - 0.65×10⁶</td>
<td>-0.25*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Benzene</td>
<td>Leukemia</td>
<td>-</td>
<td>-</td>
<td>-0.22*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,3-butadiene</td>
<td>Leukemia</td>
<td>-</td>
<td>-</td>
<td>-0.29*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Carbon tetrachloride</td>
<td>Adrenal Tumors</td>
<td>0.71×10* (0.14×10⁶)</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chromium VI</td>
<td>Lung Cancer</td>
<td>0.27×10* (0)</td>
<td>-0.88×10⁶ - 0.33×10⁶</td>
<td>-0.12*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Formaldehyde</td>
<td>Nose Cancer</td>
<td>-</td>
<td>-</td>
<td>-0.33*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Naphthalene</td>
<td>Nose Cancer</td>
<td>-0.01×10⁶</td>
<td>-</td>
<td>-0.28*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PAHPOM</td>
<td>Lung Cancer</td>
<td>0.43×10× (0)</td>
<td>-0.24×10⁶ - 0.40×10⁶</td>
<td>0.24*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tetrachloroethylene</td>
<td>Liver Cancer</td>
<td>-0.80×10* (0.01×10⁶)</td>
<td>-</td>
<td>-0.36*</td>
</tr>
</tbody>
</table>

* Significant at 1% confidence interval

** The lung cancer RR and the cancer risk due to 10 air toxics are at the level of census tract.

Among the 3 air toxics related to lung cancer, only PAHPOM had consistent positive correlation relationship with lung cancer. We then chose PAHPOM in the following case study for spatial trend of GWR coefficient.

The model comparison results were shown in Table 5. The GWR was the overall best-fitting regression model in terms of adjusted R-squared as well as the goodness-of-fit AICc statistic. The adjusted R-squared value represented the percentage of local deviance explained. In general, the higher the R-squared, the better the model fits your data. In terms of the AICc statistic, the model with the lowest AICc value is the model with the best fit. Based on this criterion, GWR regression had a slightly better fit than the OLS regression. For lung cancer, the adjusted R-squared increased from 0.18 (OLS as a non-spatial model) to 0.25 (GWR as a spatial model), which suggests a stronger correlation when information from the surrounding areas was taken into account in the spatial regression model.

Table 5. Comparison of Non-Spatial Regression Model (OLS) and Spatial Regression Models (GWR).

<table>
<thead>
<tr>
<th>Model</th>
<th>Adjusted R-Squared</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stomach Cancer RR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.06</td>
<td>49,726</td>
</tr>
<tr>
<td>GWR</td>
<td>0.07</td>
<td>49,632</td>
</tr>
<tr>
<td>Breast Cancer RR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.05</td>
<td>20,925</td>
</tr>
<tr>
<td>GWR</td>
<td>0.06</td>
<td>20,755</td>
</tr>
<tr>
<td>Lung Cancer RR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.18</td>
<td>4682</td>
</tr>
<tr>
<td>GWR</td>
<td>0.25</td>
<td>4356</td>
</tr>
</tbody>
</table>

Most GWR coefficients fluctuated widely across the research area and exceeded the confidence intervals found in the OLS regression (Tables 3 and 4), indicating a more or less varying relation between cancer RR and spatial indicators across research area. In the following section, we then had a closer examination for the spatial trends of local coefficients through case studies.

**Case Study 1: Stomach Cancer RR and Demographic Indicators**

For stomach cancer RR, the local coefficients of Asian population proportion varied across New York State and were mostly positive (84.4%), indicating a positive association for most part of New York state (Figure 1). The positive association was strongest in the southwest and northeast corner of upper state.
The association also displayed strong regional and intra-urban differences. All the elevated clusters were located at the New York City where the Asian proportion generally had a positive impact. The deficit stomach cancer clusters (with green boundary in Figure 1), on the other hand, mainly appeared at the rural area (areas with line fill symbol in Figure 1) and the coefficient values had a strong local differences within the cluster boundaries.

![Coefficient of Asian Population Proportion for Stomach Cancer RR](image)

**Figure 1.** The choropleth map that visualizes Geographically Weighted Regression (GWR) local coefficient of Asian population for stomach cancer Relative Risk (RR).

**Case Study 2: Breast Cancer RR and Socio-economic Indicators**

As shown in Figure 2 for breast cancer RR, the local coefficients of low education proportion were overall negative (98.6%), indicating that lower education was associated with a lower breast cancer risk. The small portion of positive coefficient was mainly located at the adjacent area between Brooklyn and Queens Boroughs in New York City.

Most of the highlighted clusters, either elevated clusters or deficit clusters, were located at urban areas in the downstate New York. While all the elevated clusters were located at three urban areas throughout the New York state, all the deficit clusters gathered in the New York City area.

**Case Study 3: Lung Cancer RR and Environmental Indicators**

As shown in Figure 3 for lung cancer, the coefficient values of PAHPOM were mostly positive throughout the New York State, which is consistent with the well-established relationship between PAHPOM and lung cancer. One exception was the area of long island where had an overall negative correlation to PAHPOM.
In general, areas with higher than expected lung cancer incidence (elevated clusters) were located in upstate New York and areas with lower than expected incidence (deficit clusters) were located in downstate New York. The only exception was the area of Long Island where two elevated clusters were located.

Most of the highlighted clusters, either elevated clusters or deficit clusters, were located at urban areas, with a small portion of elevated clusters appeared at the rural areas. Such inconsistent relationships may require further research for the potential driving factors.

Figure 2. The choropleth map that visualizes Geographically Weighted Regression (GWR) local coefficient of low education population proportion for breast cancer Relative Risk (RR).

Discussion

This study linked the New York Cancer Mapping Dataset with spatial exposure data sources, thus supporting spatial correlation analyses that draw consistent results with previous work about cancer risks in relation to spatially varying risk factors. In addition, we provided three case studies that mapped the varying values of the local coefficient with improved resolution, which is useful for refined evaluation into spatial trends between cancer risk and spatial indicators.

While our ultimate goal is to provide integrative spatial analytics at fine grained spatial resolution, for this initial work, we focused on the analysis of cancer counts, without patient level demographic details such as gender, age or age ranges, race and ethnicity groups. While the expected case counts capture certain demographic structure of the
population, it does not satisfy age-specific or race-specific analysis\textsuperscript{9}. The data also lack temporal information such as patient admission or discharge date for discovery of temporal patterns.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{The choropleth map that visualizes Geographically Weighted Regression (GWR) local coefficient of PAHPOM for lung cancer Relative Risk (RR).}
\end{figure}

In our future work, we will take advantage of data from New York Statewide Planning and Research Cooperative System (SPARCS), which comes with fine grained spatial information. SPARCS data comes with patient level details on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient\textsuperscript{2}. Such data also provide street level location information for each patient and healthcare facility site.

After geocoding and de-identifying addresses into census block group identifiers, our framework for integrative spatial data analytics will provide spatial queries based on coordinates or boundaries, thus linking and integrating the health records with spatial exposure data at multiple resolutions. We will first provide multi-dimensional analysis by grouping patients according to their demographic or socio-economic attributes. We will then study potential spatial clusters of disease distributions and correlations between disease risk and spatial impact factors. For example, we are interested in exploring potential hotspots of Hepatitis C or potential environment and weather factors that may have correlations with asthma.

Conclusions

Vast amounts of spatial big data are being increasingly generated and provided in the public health domain. Integrating multiple sources of spatial big data could provide new insights and create new forms of value at much higher spatial
resolutions to support community or neighborhood level public health studies. In this paper, we present our initial work on integrative spatial data analytics combining cancer incidence data in New York State, Census data and air toxics data. We focus on three representative case studies: correlation between stomach cancer and ethnicity groups, correlation between breast cancer and high education population, and correlation between lung cancer and air toxics. Our results not only are consistent with traditional studies, but also provide much refined results with improved spatial resolution. Our methods are generic and will be applied to New York State SPARCS data in the future.

Acknowledgments
This work is supported in part by NSF ACI 1443054, by NSF IIS 1350885 and by NSF IIP1069147.

References

400
Classification Models for Pulmonary Function using Motion Analysis from Phone Sensors

Qian Cheng, MS\textsuperscript{1ad}, Joshua Juen, MS\textsuperscript{1bd}, Shashi Bellam, MD\textsuperscript{2a}, Nicholas Fulara, MS\textsuperscript{2b} Deanna Close, RN\textsuperscript{2b}, Jonathan C. Silverstein, MD\textsuperscript{2c}, Bruce Schatz, PhD\textsuperscript{1acd}

\textsuperscript{1}University of Illinois at Urbana-Champaign, Urbana, Illinois USA
\textsuperscript{a}Department of Computer Science, \textsuperscript{b}Department of Electrical & Computer Engineering, \textsuperscript{c}Department of Medical Information Science, \textsuperscript{d}Institute for Genomic Biology
\textsuperscript{2}NorthShore University HealthSystem, Evanston, Illinois USA
\textsuperscript{a}Department of Medicine, \textsuperscript{b}Department of Respiratory Therapy, \textsuperscript{c}Center for Biomedical Research Informatics

Abstract

Smartphones are ubiquitous, but it is unknown what physiological functions can be monitored at clinical quality. Pulmonary function is a standard measure of health status for cardiopulmonary patients. We have shown phone sensors can accurately measure walking patterns. Here we show that improved classification models can accurately measure pulmonary function, with sole inputs being sensor data from carried phones. Twenty-four cardiopulmonary patients performed six minute walk tests in pulmonary rehabilitation at a regional hospital. They carried smartphones running custom software recording phone motion. For every patient, every ten-second interval was correctly computed. The trained model perfectly computed the GOLD level 1/2/3, which is a standard categorization of pulmonary function as measured by spirometry. These results are encouraging towards field trials with passive monitors always running in the background. We expect patients can simply carry their phones during daily living, while supporting automatic computation of pulmonary function for health monitoring.

Keywords: knowledge representation and information modeling, mobile health (patients), chronic care management (clinicians).

Introduction

A revolution in health monitoring is coming, due to widespread mobile devices. Individual measurement can generate population cohorts of similar patients with similar status, so treatments can be effectively and efficiently targeted towards all groups\textsuperscript{1}.

Mobile phones are nearly ubiquitous in the United States, with the Pew Internet Project showing 91\% ownership in May 2013, including 56\% with smartphones. Even seniors over 65 years of age have 76\% penetration of mobile phones\textsuperscript{2}. Since hundreds of millions of patients are already carrying phones, the opportunity appears for passive monitoring without adherence difficulties. We seek clinically valid physiological measures, vital signs which can be accurately monitored with smart phones.

Of the many measures that could be measured, the most important for diagnostic purposes are functional status. This is particularly true of cardiopulmonary disease, the major cause for chronic conditions in senior patients. As the patients age, their hearts and lungs slow down, and are not able to keep up with the demands. The physiological tests and measures clearly demonstrate their response to activity demand, e.g. as a patient walks, they slow down or move unsteadily when their heart and lungs cannot provide sufficient oxygen delivery to match increased exertion.

Pulmonary function is measured with a medical device called a spirometer. The patient breathes into this device, and the amount of air respired is recorded. The volume of air exhaled has been calibrated to provide standard measures of pulmonary function as discussed below. We show that a classification model can perfectly compute pulmonary function. In testing with cardiopulmonary patients, we further show that adequate inputs are the motion sensors already contained in ordinary smartphones. Simply carrying phones in daily living can measure health status.

For chronic heart and lung conditions, walk tests are widely used to assess the severity of the disease, including measures with accelerometer sensors\textsuperscript{3,4}. The Six-Minute Walk Test (6MWT) is a standard assessment\textsuperscript{5} for Chronic Obstructive Pulmonary Disease (COPD) and Congestive Heart Failure (CHF), which affects tens of millions of patients. A 6MWT measures the distance walked in six minutes back and forth over a fixed length walk way.
Normal gait requires many systems, including strength, sensation, and coordination, function in an integrated fashion, so abnormal gait is a diagnostic of many conditions\(^6\). Note that gait is the total walking pattern, the complete body motion, including swaying as well as stepping. So effective motion analysis must include more features than merely step counting, with adequate models.

Previously, we used the 6MWT to show motion features can be measured with smart phones carried by chronic patients, as sole input to a trained model that accurately predicts computes gait speed\(^7\) and oxygen saturation\(^8\), among other physiological measures for health status of chronic conditions. This study extends these results to computing pulmonary function, based on characteristic motions of health status. We also developed a phone app called MoveSense, and showed that it can record walking using phone sensors with similar accuracy to medical devices and higher accuracy than fitness devices\(^9\). This is true while running on the least expensive smart phones, such as the LG Optimus Zone, which now cost less than the portable fitness devices, such as the FitBit Flex.

**Subjects and Data Collection**

We recruited twenty-seven pulmonary patients at NorthShore University Health System, under IRB approval start November 2014. All chronic patients going through pulmonary rehabilitation in Respiratory Therapy in Evanston Hospital are now offered the option of participating in our study. Such patients are given our provided smartphones for recording their motion during a standard six-minute walk test during each visit, carried in a fanny pack. The six-minute walk tests are performed on a thirty-meter straight walkway in hospital corridor, with cones at the terminals. The patient walks back and forth on the walkway under the supervision of nurses for six minutes and the distance is recorded. The patient is permitted to stop and rest anytime, although the clock keeps going. Our software does not include this stationary data. All six-minute walk tests follow the ATS guidelines in a completely standard way\(^5\).

Pulmonary function tests are performed with a spirometer in clinical conditions\(^10\). In obstructive diseases, such as COPD, the ratio of forced expiratory volume in one second (FEV\(_1\)) as compared to an age, gender, race, and height adjusted expected value is used as a sufficient indicator to measure severity level of the disease, called predicted FEV\(_1\)% . The Global Initiative for Chronic Obstructive Lung Disease (GOLD) defines a standard for cardiopulmonary severity levels based on the predicted FEV\(_1\)% values\(^11\). There are four GOLD stages: GOLD 1 (mild), GOLD 2 (moderate), GOLD 3 (severe) and GOLD 4 (more severe). In this study, we have two GOLD 1 patients (predicted FEV\(_1\)% \(\geq 80\)), thirteen GOLD 2 patients (predicted FEV\(_1\)%: 50-79) and nine GOLD 3 patients (predicted FEV\(_1\)%: 30-49). GOLD 4 patients (predicted FEV\(_1\)% <30) typically cannot complete walk tests.

The demographic information for each group is shown in Table 1. There were two patients, Patients 12 and 24, who did not have pulmonary function tests in the record, so we eliminated these in our dataset. Patient 2 did not perform a six-minute walk test, which means there is no eligible walk data collected for this patient. This leaves twenty-four patients with walking data eligible for analysis with the models.

**Table 1.** Demographic information of each group of patients by GOLD levels. Age, height and weight are in average (minimum - maximum) format. These demographics plus gender are used to adjust the model.

<table>
<thead>
<tr>
<th></th>
<th>GOLD1</th>
<th>GOLD2</th>
<th>GOLD3</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients (Female)</td>
<td>2 (1)</td>
<td>13 (5)</td>
<td>9 (3)</td>
<td>24 (9)</td>
</tr>
<tr>
<td>Age [year]</td>
<td>69 (65-73)</td>
<td>80(67-95)</td>
<td>72 (55-85)</td>
<td>76 (55-95)</td>
</tr>
<tr>
<td>Height [m]</td>
<td>1.68 (1.55-1.80)</td>
<td>1.66 (1.24-1.83)</td>
<td>1.69 (1.55-1.83)</td>
<td>1.67 (1.24-1.83)</td>
</tr>
<tr>
<td>Weight [kg]</td>
<td>109.5 (93.0-126.1)</td>
<td>80.3(54.4-112.0)</td>
<td>81.2 (45.4-118.4)</td>
<td>83.0 (45.4-126.1)</td>
</tr>
</tbody>
</table>

Our phone app MoveSense was installed on the smartphones (a high-end Samsung Galaxy S5 and a low-end LG Optimus Zone2) for motion data collection. Our previous hardware experiment showed that the accelerometer in high-end and low-end smartphones are identical for human motion capturing, and both of them have comparable high quality as a medical accelerometer\(^8\). We take all data collected with the high-end smartphone for analysis and keep the low-end smartphone data for backup. Due to technical failure of the high-end phone, we use the data from low-end phone for three 6MWT sessions: the second session of Patient 5 and the first and the second session of Patient 16. We recorded the approximate starting time of each 6MWT session, and we let the patients stand still for seconds right before and after the 6MWT so that a pure six-minute walk can be easily extracted. The phone sensor data is downloaded into an archival database and correlated with patient medical data from the Epic EMR system.
Methodology

We established a model for GOLD stage classification. The model input is the raw three-dimensional acceleration collected by carried smartphones, and the output is a class of GOLD stage from 1 to 4 for each patient.

Preprocessing

MoveSense collects and processes the raw sensor data to extract qualified walking data. In six-minute walk test, subjects are allowed to rest during the walk test, so we must clean the dataset to detect if the patient is walking or stationary mode. We introduce a classification algorithm based on standard deviation analysis\(^\text{12}\).

We calculate the magnitude of the raw acceleration data to compress 3D signals into 1D signals, then split the signal of each session into ten-second length samples with five-second sliding window\(^\text{7}\). Rather than a static threshold of standard deviation, the classification algorithm dynamically assigns standard deviation threshold for each session by reading the whole signal thread beforehand in one-second intervals. The algorithm can detect walking or stationary mode for six-minute walk test, see Figure 1. When the standard deviation threshold is set, each one-second walking has a binary decision for walking/not-walking. Then for all ten-second samples, if the ratio of walking is larger than 0.7, we select this sample as an eligible sample based on training sets.

Feature Selection

After preprocessing, we have obtained good walking data for each patient. We compute input features for training the model by feature selection approach (FSA)\(^\text{7}\). Feature selection approach (FSA) is based on empirical knowledge. A primary parameter to measure gait is the cadence, which is the number of strides within a unit of time\(^\text{13}\). We calculated cadence by counting steps in each ten-second interval\(^\text{4}\) and dividing by length of sample.

Related studies in motion tracking by wearable devices extract a series of features from raw acceleration data\(^\text{15,16,17}\). We selected eight sufficient spatio-temporal gait parameters in both time and frequency domain. In the time domain, we selected mean and standard deviation of acceleration to describe the general distribution of the acceleration sample. In addition, mean crossing rate (MCR), root mean square (RMS), autocorrelation coefficient (AC) and coefficient of variance (CV) were calculated from time-series acceleration data. The mean crossing rate (MCR) represents the ratio of above and below acceleration. The root mean square (RMS) is a statistical measure on the variation of signal magnitude. The autocorrelation coefficient (AC) measures periodical similarity in time domain. The coefficient of variance (CV) is a normalized measure for dispersion of discrete samples. In the frequency domain, we computed the peak frequency (PF) and Shannon entropy. The peak frequency (PF) represents the frequency of peak magnitude in spectrum. Shannon entropy is an expected value of the information in the signal\(^\text{18}\).

Besides the cadence and spatio-temporal gait parameters, demographic information must also be considered in model training, just as a spirometer uses it to adjust the raw values from patient respiration. Unlike selecting demographic cohort and training models by cohort in our previous research\(^\text{7}\), we keep the four basic demographic parameters: age and sex, height and weight as factors in the input feature vector. With such input feature vector, we train universal models which ideally can be applied on general population. Overall, for each ten-second sample, the input feature vector \(x\) contains thirteen independent features -- covering stepping, moving, and demographics.

![Figure 1](image-url)
Model Training

Support vector machines (SVM) are a class of machine learning algorithms, widely used in learning classification models. We apply RBFSVM, the SVM with radial basis function kernel, to train the model for GOLD classification. SVM allows classification by remapping multi-dimensional vector $\vec{x}$ into a higher dimensional space and determining a hyperplane or a set of hyperplanes that separate different classes. The hyperplanes optimize the separation between data points in the given space, yielding the classification Equation (1). SVM classification requires a training set to find the variables $\alpha_i$ to determine $f(x)$ with a given kernel function and given inputs.

$$f(x) = \sum_{i=1}^{N} \alpha_i \cdot K(x_i, x)$$

From Equation (1) we know that different kernel functions affect the overall performance of the trained model, previously we applied simple linear kernel in model training and obtain the technical baseline. In this study we applied RBFSVM, using the radial basis function kernel to train the SVM model, as Equation (2). We set the hyper-parameter $\gamma = \frac{1}{D}$, based on the number of dimensions in training samples. $x_i$ and $x_j$ are two arbitrary sample vectors in the training set. In our dataset patient’s status are labeled as one of the three GOLD stages (GOLD 1, GOLD 2 and GOLD 3). One-against-all strategy is applied to train this three-class classification model. The tolerance of termination criterion is set as 0.001 and the insensitive-loss variable $\xi$ is set as 0.1.

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \quad \gamma = \frac{1}{D}$$

Model Validation

Since patients only take pulmonary function tests during the initial walk test (6MWT1) but not for the following test (6MWT2, 6MWT3), we only know the FEV1% predicted values for their first 6MWT session. We label samples during the first 6MWT with corresponding GOLD stages. Thus the dataset is partially labeled. We then select all labeled samples to form the training set. To avoid over-fitting, 10-fold cross validation is applied for self-validating the classification model. That is, randomly split the training set into ten parts, and each time select one out of the ten subsets as the test set and the other nine to train the model. After ten folds of model training and testing, we obtain predicted results of all samples in the dataset with a model trained by other samples from the same cohorts.

Voting Status with Thresholding

A classification model obtains GOLD stages for each ten-second walking samples. However, each six-minute walk test contains multiple ten-second walking samples, up to 37 if no stopping during the test. For cardiopulmonary patients, the severity will not change within a single walk test. But a walking sample may be affected by other factors and lead to false classification. Majority voting is robust rule to obtain single decision from multiple samples. Majority voting worked in our previous studies so we apply to obtain predicted GOLD stages for each patient. Based on empirical observations, we apply a strict threshold of 85% to accept the voting, which means only when the majority group gets more than 85% of the votes, the decision will be accepted. Otherwise, if none of the groups obtains higher than 85% votes, we mark the status as pending and clinically analyze the risk of this walking session.

Results

The dataset contains 1204 ten-second walking samples (39 six-minute walk tests) from 24 different patients, including 769 labeled samples from the first 6MWT session of each patient. We train classification models only using all labeled samples. Ten-fold cross validation is applied to self-validate the model. We then utilize the trained model to predict all unlabeled samples.

Evaluating the model

Besides the RBFSVM model, we train a classification model as in our previous study, which applies SVM with linear kernel as baseline. The results are shown in Table 2. The overall accuracy of the RBFSVM model is 98.60% (100% for GOLD 1, 99.28% for GOLD 2 and 98.37% for GOLD 3). While the overall accuracy of the linear SVM model is 86.35% (100% for GOLD 1, 88.31% for GOLD 2 and 82.02% for GOLD 3). The RBFSVM model significantly improves classification accuracy for GOLD 2 and GOLD 3 samples from 82% to 98%.
Table 2. Self-validation results for RBFSVM model and linear SVM model. We compare the accuracy of classification, which is all correct classified samples divided by the total number of samples. This is by ten-second interval, not by patient. See Figure 2 for patient-level results.

<table>
<thead>
<tr>
<th></th>
<th>RBF SVM</th>
<th>Linear SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>98.60%</td>
<td>86.35%</td>
</tr>
<tr>
<td>GOLD1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>GOLD2</td>
<td>99.28%</td>
<td>88.31%</td>
</tr>
<tr>
<td>GOLD3</td>
<td>98.37%</td>
<td>82.02%</td>
</tr>
</tbody>
</table>

Figure 2. Evaluation of GOLD classification for labeled 6MWT sessions. The bar-plot shows the portion of predicted GOLD stages for all labeled walking samples of each patient. Except for Patient 7, 9 and 17, all other patients are 100% correctly predicted within every interval. After voting, all patients are predicted as their actual GOLD stages.

After obtaining the classification results for ten-second samples, majority voting is applied for categorizing each patient into a target GOLD stage. We perfectly categorize all patients into the corresponding GOLD stages summarized from their FEV1% predicted values, shown in Figure 2.

Classifying Unlabeled Samples

We train a universal classification model with all 769 labeled samples and apply this model to classify all unlabeled samples. There are 435 unlabeled samples from thirteen unique 6MWT sessions from ten patients, where three patients have third 6MWT sessions (5,16,21). Two out of the thirteen walking sessions are classified differently.
from the patients’ first labeled 6MWT session. For Patient 1, the second 6MWT session is 100% predicted as GOLD 2 while the first 6MWT shows the patient is in GOLD 3, as the FEV1 indicates with spirometry. For Patient 16, the second 6MWT session is 68.97% predicted as GOLD 2 and 31.03% predicted as GOLD 3, while the first 6MWT session of this patient is labeled as GOLD 3, as the FEV1 indicates with spirometry. Referring to the rule of voting with thresholding, none of the predicted GOLD states reaches the 85% threshold so the second session of this patient is in pending status. Patient 16 also has a third 6MWT, which is 10.34% predicted as GOLD 1, 3.45% predicted as GOLD 2 and 86.21% predicted as GOLD 3. So the third session of Patient 16 is decided as GOLD 3, the same as the first session. For all other patients, the classifications of the unlabeled 6MWT sessions are 100% identical to their first 6MWT sessions for all intervals. The GOLD classification for each 6MWT session is shown in Figure 3.

**Figure 3.** Evaluation of GOLD classification for unlabeled 6MWT sessions. The bar-plot shows the portion of predicted GOLD stages (PGOLD 1, PGOLD 2, PGOLD 3) for all unlabeled walking samples of each patient. Except for Patient 1 Session 2, Patient 16 Session 2 and Session 3, all other patients are 100% predicted as the GOLD stages they are labeled by the initial 6MWT session (Session 1).

### Sufficient Walk Length for Accurate Classification

The main results above show that the classification model yields high accuracy when the model is trained with full six-minute walk for each patient session. Additionally, we analyze the effect of walk length in training progress to classification accuracy of the model. We randomly select different lengths of walking in the 6MWT session for each patient, from 30 seconds to six minutes (30s, 40s, 50s, 1min, 2min, 4min, 6min). We train and evaluate the model on the selected walking periods using the same strategy as above to obtain the predicting accuracy for each length. The result is shown in Figure 4. When the walk length is 120 seconds (two minutes), the predicting accuracy has reached 100% and remains there through the full six minutes. Note it is nearly perfect after just one-minute. So we assume a two-minute good walking for each patient is sufficient to train a classification model for detecting GOLD status. Here “good walking” refers to the steady pattern, as during a walk test, where high accuracy of the model is achieved, since the patient changes speeds reflecting their pulmonary function.
Figure 4. Classification Accuracy with Different Walk Length in Training. The accuracy in ten-fold cross validation reaches ~100% after two-minute (120-second) length.

Discussion

High Accuracy Classification on Labeled Walking Sessions

There were twenty-four patients participating in at least the first six-minute walk test session. The model does perfect computation for their pulmonary function. That is all patients are categorized into the correct GOLD stage. In this case the pulmonary function tests are performed the same day as the six-minute walk test, so we assume all patients’ predicted FEV1% represents their current severity of respiratory limitation when we collect their motion data and the phone motion could fairly match the lung function. Figure 2 summarizes this perfect computation of health status for pulmonary function.

In more detail, for twenty-one patients, all their ten-second walking samples are identically predicted, and correctly reflect the patient’s GOLD stage. This indicates spatio-temporal motion reflects more stable information than the one-dimensional walking speed. Patient may speed up or slow down during the six-minute walk test, but the spatio-temporal motion will always yield the correct health status, by measuring stability deviation of walking movements. Based on this outcome, we assume that six-minute walk tests are not necessary for detecting health status of cardiopulmonary patients. Instead, if we can find any good walking sessions that are identical to the walking during the six-minute walk test, spatio-temporal motion data from such walking session can be used to compute GOLD status. As Figure 4 shows, our model only needs 2 minutes of good walking for perfect computation.

For three patients, not all ten-second samples are classified correctly, including Patient 7, Patient 9 and Patient 17. For Patient 9, one ten-second sample out of thirty-three eligible samples gets false classification, while for Patient 17, one ten-second sample out of twenty-five eligible samples gets false classification. These can be simple statistical error so we can ignore them. For Patient 7, five samples out of thirty-seven eligible samples get false classification, which takes 13.5% of the overall classification classifying GOLD 2 for a patient measured as GOLD 3. So voting with 85% threshold would still predict correct status compared to spirometry that measures FEV1%.

In addition, for Patient 7, the spirometry measurement was 5 months earlier than the 6MWT, whereas nearly all other patients had their PFT within 1 month. This delay could lead to false classification, due to change in clinical status. For this patient, they also had a medication change to begin COPD inhalers during this period, so we expect an improved status with better pulmonary function. Thus there is a clinical explanation for lower model inaccuracy.

Identical Classification on Unlabeled Following Walking Sessions

From medical literature, six-minute walk distance (6MWD) is considered as a gold standard for diagnosis of COPD severity and other cardiopulmonary health status, but it is more accurate for severe COPD than moderate or mild COPD. In our experiment, ten patients performed at least two six-minute walk test sessions, as shown in Figure 3, usually spaced several weeks apart. We compare their initial 6MWDs and following 6MWDs, shown in the Bland-Altman plot in Figure 5.
There are 13 sessions, since 3 patients also did a session 3, but only 11 points since we eliminate 2 because no actual 6MWD recorded (Patient 5 Session 3 and Patient 16 Session 3). The initial 6MWDs are in average of 310.8m (±95.16m) and the following 6MWDs are in average of 367.1m (±106.5m). Generally, the patients walked further (had greater 6MWD) in later sessions, as expected for rehabilitation patients.

Our analysis indicates that 6MWD is not sufficient for measuring COPD severity levels, certainly compared to pulmonary function for cardiopulmonary patients. In our dataset, the GOLD 2 and GOLD 3 patients are not well distinguished by 6MWD. The 6MWD of GOLD 3 patients can be as high as 350 meters but the 6MWD of GOLD 2 patients can be below 250 meters. Note the variation between two different 6MWDs of a patient becomes larger when the distance is shorter, e.g. the second 6MWD of one GOLD 3 patient (Patient 23) is 150 meters longer than the initial 6MWD, even though performed just twenty days after he began the rehabilitation process.

Our classification model again had high accuracy, since such variation on total distance does not affect the stability of patients' motion pattern. Eight out of ten patients are detected strictly as the same GOLD status for their second or third 6MWT sessions. All of the ten-second walking samples of their following 6MWT sessions are identical to their original GOLD status, even for Patient 23, the one with the largest variation of 6MWDs. This “unfair” 8 of 10 model accuracy for later sessions compared to the initial GOLD status is shown in Figure 3.

The two patients who are not correctly classified at subsequent sessions are Patient 1 and Patient 16. For Patient 1, GOLD 2 is classified at the subsequent session but the PFT measure indicates GOLD 3. Patient 1 had a clinical diagnosis of both COPD and CHF when he was enrolled in cardiac rehabilitation. His medication regimen for CHF was adjusted at the start of his rehabilitation program and his motion was noted to improve, likely from better control of his CHF, so in fact this “unfair” comparison of initial GOLD status to improved GOLD status with the model may be measuring actual improvement rather than model inaccuracy. Congestive Heart Failure is less closely correlated to pulmonary function than Chronic Obstructive Pulmonary Disease, limiting the usefulness of GOLD criteria for defining the severity of functional limitation in CHF. The second 6MWD of Patient 1 is 72.5 meters longer than the initial 6MWD.

Monitoring Severity Change Longitudinally

Patient 16 had diagnosis of sarcoidosis, which is a lung disease that has variable effects on pulmonary function. Sarcoidosis cannot be well categorized by disease severity based upon GOLD criteria, as the physiologic limitations are less clearly correlated with predicted FEV1%, as in COPD. However, this patient was measured for pulmonary function via spirometry for FEV1 indicating GOLD 3, which was correctly classified in Session 1.

Patient 16 performed three 6MWT sessions at different visits, among which the second session was about one month later than the initial session and the third session is about five month away from the initial session, which is beyond the three-month rehab period. In Session 2, there was mixed classification of 2/3 intervals indicating GOLD 2 and 1/3 indicating GOLD 3. The patient was likely within transition between status levels, due to rehabilitation progress. During Session 3, not only was it much later, but there was a clinical note that the patient had suffered a significant illness which was the reason to return to rehabilitation. This session was the only true mixture, with some classifications of GOLD 1/2/3, although 86% of the intervals indicated GOLD 3 once again. So the lung
physiological function likely worsened and the patient regressed back to the original status level. Our model would have predicted this correctly, if the threshold mentioned earlier of 85% was implemented. Thus there is clinical evidence that our model can predict changes in pulmonary function.

**Conclusion**

In this study, we expand our previous results into accurate computation of pulmonary function with universal models. This requires utilizing demographic features with better training sets (ATS standard) and better statistical models (RBF-SVM), which utilize characteristic motions in addition to step counting. Every patient now has correct modeling of GOLD level, even sample by sample (10 seconds) not only walktest by walktest (6 minutes). Measuring motion is a potential solution for passive monitoring, which distinguishes this work from the phone applications measuring pulmonary functions with the microphone. A passive monitor has compliance advantages over active phone applications. That is, the patient simply uses their personal phone as usual during daily activities, no special actions or special experts are necessary. The microphone “spirometer” also has adherence difficulty, the patient must breathe properly into the proper place, also a difficulty with a medical spirometer.

The trained model is providing perfect computation of pulmonary function category (classifying GOLD stage). This is especially true for the first session with pulmonary patients, where spirometry has just measured the pulmonary function so the pulmonary function value can be directly compared to the motion value. For senior patients undergoing pulmonary rehabilitation, respiration and motion are closely correlated. This is why the classification is so accurate, especially in the hospital setting with monitored walk test. The result is encouraging, so we shall be relaxing environmental constraints on our measurements in forthcoming studies, moving closer to daily activities in the real world of cardiopulmonary patients. Under expanded IRB, we now allow selected patients to bring smart phones home from rehab, to measure their motions and predict their pulmonary functions during daily activities.

The home trials are more complex than the hospital trials thus far. The patient will be doing many different activities throughout the day, while the periods when they are carrying their phones could all be potentially recorded by a passive monitor that is always running. The passive monitor phone app used in the home is a re-implementation of the active monitor used in the clinic, which only records when the phone is in steady motion and only archives when the patient is good walking. Since the models have high accuracy during walk tests, we are developing software for activity recognition. This is a well-studied problem in computer science, for using motion sensors to detect which activity a person is performing. General activity recognition is hard, but our problem is a specialized version of that. The recognition filter in the phone app need only detect good walking, when the body motion is similar to a walk test. This paper reports on initial progress of autodetecting stopping during walktests. Since the amount of walking needed for accurate computation is very small, only 1 or 2 minutes per session, the filter can be tuned up, for high precision only during periods when there is definite measurement of good walking.

We are optimistic that full-scale clinical trials are possible to compute health status for cardiopulmonary patients. They need only carry their own phones and do some limited good walking during the day when the passive monitor is recording unobtrusively. Since the cheapest smartphones can accurately support this computation, large-scale population measurement can become an everyday reality for health systems.

**Acknowledgements**

At the University of Illinois at Urbana-Champaign, the Institute for Genomic Biology provided facilities for software development and data analysis. At NorthShore University HealthSystem, the Department of Respiratory Therapy at Evanston Hospital carried out patient testing and phone recording, with co-author Deanna Close, RN, as supervising pulmonary nurse. All clinical experiments were protected by NorthShore IRB EH15-025 entitled “Mobile phone software MoveSense versus traditional exercise testing to assess health status”, with co-author Shashi Bellam, MD, as supervising pulmonary physician. We acknowledge Brian Edwards and Catherine Zhu for managing and anonymizing phone data and medical records inside NorthShore University HealthSystem using the REDCap archiving system.
References

Usability Testing of a Web-Based Decision Aid for Breast Cancer Risk Assessment Among Multi-Ethnic Women

Austin M. Coe, MPH1, William Ueng, MPH1, Jennifer M. Vargas, BS1, Raven David, MPH1, Alejandro Vanegas, MS1, Katherine Infante, MPH1, Meghna Trivedi, MD1, Haeseung Yi, MPA1, Jill Dimond, PhD2, Katherine D. Crew, MD, MS1, Rita Kukafka, DrPH, MA1

1Columbia University, New York, NY; 2Sassafras Tech Collective, Ann Arbor, MI;

Abstract

Chemoprevention with antiestrogens could decrease the incidence of invasive breast cancer but uptake has been low among high-risk women in the United States. We have designed a web-based patient-facing decision aid, called RealRisks, to inform high-risk women about the risks and benefits of chemoprevention and facilitate shared decision-making with their primary care provider. We conducted two rounds of usability testing to determine how subjects engaged with and understood the information in RealRisks. A total of 7 English-speaking and 4 Spanish-speaking subjects completed testing. Using surveys, think-aloud protocols, and subject recordings, we identified several themes relating to the usability of RealRisks, specifically in the content, ease of use, and navigability of the application. By conducting studies in two languages with a diverse multi-ethnic population, we were able to implement interface changes to make RealRisks accessible to users with varying health literacy and acculturation.

Introduction

Breast cancer is the most commonly diagnosed and second deadliest cancer among women in the United States.1,2 Based upon randomized controlled trials, chemoprevention with antiestrogens, specifically tamoxifen, raloxifene, anastrozole, and exemestane, can reduce the incidence of breast cancer among high-risk women by 30-70%.3–6 It is estimated that at least 15% of women in the United States meet high-risk criteria for breast cancer, defined as a 5-year risk greater than 1.67% or a lifetime risk greater than 20%, according to the Gail model.7 However, of these women fewer than 5% who are offered chemoprevention therapy reported taking the medication.8 While this low uptake is partially attributed to concerns about side effects, lack of knowledge about chemoprevention among both patients and primary care providers (PCPs) is another a major contributor to low uptake.9

Recognizing and making decisions about health can be a difficult task for patients, particularly when there are multiple options for care and patient preferences are not well defined.10 The United States Preventive Services Task Force (USPSTF) recommends that physicians engage in shared decision-making with patients regarding chemoprevention and screening for breast cancer.11 Studies have indicated that patient decision aids (PtDAs) are an important part of influencing a woman’s decision to seek certain types of treatment for breast cancer.12 PtDAs for considering breast cancer screening options and prevention strategies have been developed to assist women with assessing their risk and making informed choices about their health.13–17 While PtDAs for screening were successful in reducing over-screening rates, chemoprevention uptake among high-risk women exposed to a PtDA on breast cancer chemoprevention remained low up to 3 months after use.15–17 PtDAs that provide information that is tailored to the risk status of the patient tend to be more efficacious than those that do not.10

We have developed an innovative web-based PtDA, RealRisks, to provide high-risk women with more information about breast cancer chemoprevention. RealRisks is designed to improve users’ accuracy of risk perception while presenting them with information about breast cancer risk and chemoprevention options that can be tailored to their preferred method of receiving health information. The PtDA offers users the ability to input personal information in order to engage with modules that reflect the prevention options that are available for women with the same risk level. Patients who complete RealRisks are provided with an action plan for discussing prevention options with their PCP. RealRisks is designed to be used in conjunction with the Breast cancer risk NAVigation (BNAV) tool, a clinical decision support tool for PCPs that is integrated into the electronic health record.18 Together RealRisks and
BNAV provide comprehensive information to high-risk women and their PCPs in order to promote shared decision-making about chemoprevention.

We previously conducted focus groups and participatory workshops with an early prototype of RealRisks, which specifically targets women from diverse racial/ethnic and educational backgrounds. In particular, we demonstrated that after exposure to RealRisks, women with low or high numeracy levels had improved breast cancer risk perceptions and the majority found the tool useful and easy to use.19 The aim of this study was to assess the usability of a more advanced prototype of RealRisks, which is available in English and Spanish, by evaluating issues related to the design, content, and user engagement of the PtDA. We sought to make improvements to the application that would make RealRisks understandable to multi-ethnic women with a broad range of health literacy and acculturation levels.

Methods

Participants and Recruitment

Participants in the study were recruited from a database of women who had undergone routine screening mammography and previously consented to be contacted for future research. Participants were contacted by telephone and offered morning and afternoon sessions to partake in usability testing. Subjects who indicated that they were unable to use a computer or had serious issues relating to reading a computer screen were not eligible for the study.

We initially recruited 5 English-speaking subjects for the first round of usability testing. For the second round of testing, we recruited 2 English-speaking and 4 Spanish-speaking subjects after changes were made to the program. The literature has shown that conducting usability testing with at least 5 subjects will usually identify upwards of 80% of design issues, and that 10 subjects will identify a minimum of 82% of usability concerns before becoming repetitive.20,21 In total, 11 subjects participated in the usability testing for RealRisks. All subjects signed written informed consent and the study was approved by the Institutional Review Board at the Columbia University Medical Center (CUMC).

The RealRisks Decision Aid

The RealRisks PtDA is tailored to provide personalized information about chemoprevention to high-risk women in both English and Spanish. The PtDA contains 6 modules that provide various information about risk assessment and chemoprevention (Figure 1). Each module includes a comic following a fictional character, Rose, who discusses breast cancer risk with family, friends, and health care providers (the light narrative), a series of informational slides (the dense narrative), and interactive games (Figure 2). For example, the first module introduces users to the concept of risk through a “clicking” game that demonstrates the 5-year and lifetime breast cancer risk for an average 50-year old woman using a pictograph with 100 clickable women (Figure 4). This experience-based risk interface asks players to continue to click (e.g., sample from a population of women) to better learn the meaning of a given pre-set probability (i.e., 12 out of 100 or 12% of women will develop breast cancer in their lifetime) and has been previously shown to improve risk perception in high and low numeracy women.19,22 The second module outlines various risk factors for breast cancer and prompts the participant to calculate their Gail risk score and allows users to revisit the previous risk game using their personal risk for breast cancer. The third module asks users to input their family tree, with special attention to family members with a history of cancer. The fourth and fifth modules are tailored to the woman’s level of breast cancer risk and includes preference elicitation games (Figure 3). Module four gives information about BRCA genetic testing, whereas module five describes different breast cancer chemoprevention options. Preference elicitation is done through the use of scales that allow users to rate how important the risk of developing breast cancer, risk of side effects with chemoprevention, and various personal factors are to their decision about taking chemoprevention. The sixth module opens for all participants and provides information about lifestyle behaviors than can lower the risk for breast cancer. Completing all six modules generates an action plan that includes all of their inputted information that can be printed and shared with their PCP during their next clinic visit. A copy of the action plan is also provided to the PCP through the electronic health record via the BNAV tool. RealRisks is designed to be accessible on a computer, tablet, or smartphone. Development of the risk clicking game and other components of RealRisks have been addressed in previous studies.18,19
Usability Testing

Usability studies were conducted in English and Spanish by pairs of researchers who took on the role of moderator and observer and were bilingual for the Spanish-speaking participants. The moderator was tasked with receiving consent from the study subjects, orienting the subject to the PtDA, and providing guidance and assistance as the participant completed the study. The observer used Morae Observer software to view a livestream of the subject and code the usability study in real time on a separate laptop. This livestream captured the audio and visual image of each subject, as well as non-verbal reactions. The moderator was positioned across from the subject, while the observer was in a corner of the study room. All usability studies took place at CUMC.
Participants consented to be audio and video recorded. Prior to beginning the usability testing, subjects were asked to complete a short survey to indicate their familiarity with computer and web-based technology and validated measures of acculturation and health literacy. Study subjects were then introduced to the RealRisks platform and instructed to create an account using predetermined patient profiles and scenarios. These preset profiles included simulated health information that would activate both the genetic testing and chemoprevention eligibility modules without requiring subjects to disclose their personal medical history to the research team. Participants were then asked to complete the six RealRisks modules in any order they desired and follow a think-aloud protocol as they moved through the PtDA. In this protocol, subjects voiced their feelings and intentions about chemoprevention, the design of the application, and their understanding of the information as it was presented to them. When subjects ceased thinking aloud as they used the PtDA, they were prompted by the Moderator with questions relating to the section of the application they were on. The most frequently used prompts included “how are you feeling about the information on this page”, “what are you currently thinking about”, and “why did you decide to move to that section”. Prompting was not necessary for the majority of study subjects. The moderator did not interact with the participant beyond prompting and answering direct questions regarding the application.

Each usability study lasted approximately two hours or until the subject completed the entire RealRisks PtDA. The PtDA was considered completed once a subject had gone through all six modules of RealRisks and viewed their action plan. At the end of the testing period, all subjects were asked to complete the System Usability Scale (SUS), a validated 10-item questionnaire used to determine the usability of a software. The results of the SUS were used to determine the current usability status of RealRisks.

Data Collection and Analysis

Morae Recorder, Observer, and Manager software version 3.3.4 (Techsmith Corporation, Okemos, MI) was used for all usability recording and analysis. Morae software allows for the creation of usability study templates to be created with Tasks and Markers for ease of coding. Markers were used to label points in the video such as “error”, “participant prompted”, and “observation”. Tasks were used to identify when subjects had completed activities in RealRisks, such as “created an account”, “finished module 1”, and “viewed action plan”. All participant recordings were coded by two researchers – the first being the Observer and the second being an additional researcher involved in the study, usually the Moderator. The first round of coding was done in real time, while the second occurred immediately after the study.

Qualitative coding was used to identify themes regarding the design, content, and user engagement with RealRisks. Recurring issues brought up by participants were sorted into whether they related to the dashboard, light narrative, dense narrative, games, or action plan. SUS scores were calculated using the standard methodology described by Brooke. The results of the usability studies were used by the research team to identify opportunities to improve RealRisks. Feedback given to the programming team was used to update RealRisks prior to implementation of the tool among high-risk women.

Results

Participants

For the first round of usability studies, a total of 5 English-speaking subjects were enrolled. Of these subjects, 1 had difficulty with reading some information on the screen because she forgot her glasses – this subject was given assistance by the moderator to read the screen. In the second round of usability studies, 2 were English-speaking subjects and 4 were Spanish-speaking subjects. A total of 11 subjects took part in the two rounds of usability testing for RealRisks. Our sample population was relatively diverse, in terms of education level, race and ethnicity, health literacy and acculturation (Table 1). The English-speaking group had a higher median health literacy than the Spanish-speaking group, however, in both cohorts health literacy ranged from low to high. The English-speaking group was more acculturated than the Spanish-speaking group. Two subjects spoke a language other than English or Spanish – the one subject who identified her additional language spoke Amharic.

All subjects owned a mobile phone, and 9 owned a smart phone. All but 3 subjects owned a computer of some type, although at least 1 of these subjects indicated that she had a computer in the house but was not the owner. Of the 11 subjects, 10 had access to the internet in their home.
Table 1. Baseline characteristics of study subjects

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>English-Speaking (N = 7, 64%)</th>
<th>Spanish-Speaking (N = 4, 36%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age, years (range)</td>
<td>57 (29-69)</td>
<td>56.5 (47-71)</td>
</tr>
<tr>
<td>Race and Ethnicity, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>1 (14)</td>
<td>0</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>5 (71)</td>
<td>0</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0</td>
<td>4 (100)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (14)</td>
<td>0</td>
</tr>
<tr>
<td>Education Level, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School or Less</td>
<td>2 (29)</td>
<td>3 (75)</td>
</tr>
<tr>
<td>Some College or Vocational</td>
<td>2 (29)</td>
<td>1 (25)</td>
</tr>
<tr>
<td>Graduate School or Professional</td>
<td>3 (43)</td>
<td>0</td>
</tr>
<tr>
<td>Health Literacy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range, 0 [low] – 4 [high])</td>
<td>3 (1.67-4)</td>
<td>2.33 (1.33-4)</td>
</tr>
<tr>
<td>Acculturation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range, 1 [low] – 5 [high])</td>
<td>5 (3-5)</td>
<td>1.67 (1-3.67)</td>
</tr>
</tbody>
</table>

Theme Identification

Usability themes were grouped into Content, Ease of Use, and Navigation, which were used to identify potential usability problems that could be addressed by improving RealRisks. All themes were marked as being concerned with the dashboard, light narrative, dense narrative, games, or action plan. Major recurring themes and corresponding changes are described in Table 2.

Table 2. Themes identified and addressed during the usability studies.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Location</th>
<th>Description</th>
<th>Changes Made</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Light Narrative</td>
<td>Comics are “talking down” and too simplistic; alternatively comics are highly relatable</td>
<td>Users can navigate through light and/or dense narratives based upon preferences for cognitive load</td>
</tr>
<tr>
<td></td>
<td>Dense Narrative</td>
<td>Cancer development and DNA/genetic information is confusing and too technical</td>
<td>Language and pictures were simplified</td>
</tr>
<tr>
<td></td>
<td>Games</td>
<td>Preference elicitation game does not have non-clinical factors, is confusing, not sensitive enough to patient preferences</td>
<td>Scales redesigned to include more preference factors, which were also summarized in the action plan</td>
</tr>
<tr>
<td></td>
<td>Action Plan</td>
<td>Confusion over who the action plan is meant for and what the subject should do with it</td>
<td>Action plan redesigned and given additional instructions</td>
</tr>
<tr>
<td>Ease of Use</td>
<td>Dense Narrative</td>
<td>Too much text to read, would prefer audio</td>
<td>Audio feature made more prominent/visible</td>
</tr>
<tr>
<td></td>
<td>Games</td>
<td>Risk assessment clicking game not giving enough feedback, took too long to complete</td>
<td>Clicking feedback improved, number of clicks needed were shortened</td>
</tr>
<tr>
<td></td>
<td>Games</td>
<td>Preference elicitation game had many bugs, drag and drop functionality was difficult for many participants</td>
<td>Scales redesigned to use sliding scales for preset preference options</td>
</tr>
<tr>
<td></td>
<td>Games</td>
<td>Family tree difficult to understand without a legend or key</td>
<td>Legend added to family tree</td>
</tr>
<tr>
<td>Navigation</td>
<td>Dashboard</td>
<td>Difficult for users to determine which module they had just completed and how far along they were in RealRisks</td>
<td>Checkbox functionality improved</td>
</tr>
<tr>
<td></td>
<td>Dashboard</td>
<td>Users asked for some sort of aid to guide them through the application</td>
<td>Tutorial videos in development to assist users with low computer literacy</td>
</tr>
</tbody>
</table>
Content Issues

Most problems arose in the dense narrative and during the genetic testing and chemoprevention preference elicitation games. Two subjects were vocal about their distaste for the light narrative, indicating that they thought the comics were “speaking down” or “not something I would read”. However, subjects who aligned more closely with the target population identified the cartoon figure of Rose as being “like me” and performing actions that were in line with what they would do in real life. Based on this feedback that some users prefer the light narrative whereas others dense, RealRisks was updated so that end users can complete either the light and/or dense narrative sections to trigger the creation of the action plan based upon their preferences for cognitive load.

Certain sections in the dense narrative were frequently confusing to subjects. Slides in depicting the growth of breast cancer cells was described as hard to understand and several subjects either asked the moderator for further information or described frustration with the use of technical words like “hyperplasia” or “atypia.” These concerns were addressed by improving the picture of breast cancer growth on this slide and using layman’s definitions in the rollovers on the slide. Similarly, use of the metric system on some slides was difficult to understand for at least two subjects. The dense narrative was edited to include measures in imperial units. A body mass index (BMI) chart that appears in the lifestyle factors module was difficult to read for subjects, with one staring at it for a period of time and expressing confusion at what she was looking at. This chart will be amended in future versions of the PtDA.

The most major content issue occurred with the preference elicitation games in modules 4 and 5 and overlapped significantly with ease of use. These games were designed to allow subjects to consider a variety of factors that may influence their decision-making about genetic testing or chemoprevention. In the original version of RealRisks, subjects were presented with a set of scales for genetic testing and each type of chemoprevention. Users would drag different factors onto the scale to indicate whether they considered it a positive or negative factor. The purpose of this activity was to elicit and visualize the user’s preferences. Content issues with this game included confusion over the purpose, lack of clarity regarding how the scale was weighted, and an inability of the scale to capture preferences outside of clinical factors such as “reduced risk of breast cancer” and “increased risk of endometrial cancer”. These concerns were addressed by completely redesigning the game to make the scale more intuitive. The changes included embedding the sliders directly into the page and allowing the scales to move with each slider. A Likert-type scale and color was also added to each slider. Examples of new nonclinical preference factors added to the game include “taking control of my health”, “don’t trust medications”, and “availability of social support”. A comparison of the before and after of this interactive game for chemoprevention can be seen in Figure 3.

Figure 3. Comparison of the original (left) and updated (right) versions of the preference elicitation scales for chemoprevention

Users were also confused about the purpose and structure of the action plan at the end of RealRisks. Original versions of the action plan did not specify explicitly that it was to be shared with her health care provider, and the layout of the page made it difficult to read. These issues were addressed by including more information about the purpose of the action plan and restructuring the page to be easier to comprehend.
Ease of Use Issues

Ease of use concerns were the most frequently brought up usability issue with RealRisks, but were generally minor to fix. These concerns primarily related to how easy it was to use some of the features of the games. In regards to the preference elicitation scales, there were issues with the sensitivity of the scales and the drag and drop functionality. We amended the tool to have all options preloaded, so subjects would not have to drag individual boxes (Figure 3).

For the risk assessment games, users were asked to click on a pictograph of 100 figures to identify one with breast cancer, which is a novel experience-based format for conveying risk information to those with low numeracy. Participants stated that the game took too long to complete and that it wasn’t clear how they were supposed to move on from the game (Figure 4). These issues were amended by shortening the number of clicks necessary to identify a breast cancer patient, improving the feedback for when a person was clicked, and adding a pop-up window to automatically find the patient after a certain number of clicks. Another issue with the clicking game that was brought up by both English and Spanish-speaking subjects was that the different rounds of the game were difficult to tell apart. Explanations of the risk of breast cancer appear in a yellow box to the right of the game, however, some participants found it confusing that the 5-year risk game was not further differentiated from the lifetime risk game. This concern was addressed by expanding the instructions and clearly labeling the two games.

Figure 4. The RealRisks “clicking game” used to visually demonstrate breast cancer risk

Some participants voiced difficulty with reading all of the information on the dense narrative screen – either it hurt their eyes or they would prefer to listen to the material. RealRisks was designed to include an audio option, and this feature was added in English and Spanish. Another ease of use issue occurred with the family tree in the family history module. In this game, subjects are asked to create a family tree by inputting information about their family. However, the final product did not include an index for what different shapes, colors, and lines meant, therefore, a legend was added to the family tree.

Navigation Issues

Navigation concerns were a minor recurring concern among study participants. Subjects generally chose to follow through the modules in order and found the process of moving through RealRisks to be intuitive. One recurring issue relating to navigation was confusion regarding where in the application the subject was supposed to go next. After completing a module, subjects would be brought back to the main screen with no indication of which module was meant to be next. This was addressed by improving the performance of a “checkbox” feature, that indicated to the subjects which items they had completed. When users start RealRisks, they are briefly guided through the features of the PtDA by a series of on-screen instruction boxes. One recommendation that was brought up by older English and Spanish-speaking subjects was the desire for some sort of tutorial to help guide participants through the rest of the PtDA. Another suggestion was to turn the introductory guide into a video to better explain the features of RealRisks. These recommendations will be implemented in a future version of the PtDA.
System Usability Scale Scoring

Every participant completed the SUS questionnaire at the end of the usability study. Scores varied appreciably between English and Spanish-speaking subjects. English-speaking users gave RealRisks a median score of 80.00 (range, 55.00-95.00), while Spanish-speaking users gave the application a median score of 66.30 (range, 55.00-75.00). The mean score between the two groups was 72.00. An empirical review of SUS scores has indicated that a SUS score between 65 and 70 is considered “ok” or “high marginal”, while a score above 70 is considered to be “good” or “acceptable” usability for a web-based application. This would indicate the need for further analysis and improvement of the RealRisks application based on the feedback we received from Spanish-speaking users.

Discussion

The aim of this series of usability studies was to evaluate RealRisks in order to better tailor the patient decision aid (PtDA) to women at high-risk for breast cancer, particularly multi-ethnic women with a broad range of health and computer literacy. By conducting studies in both English and Spanish, we were able to identify several issues across three themes that reduced the usability of the application for our target population. This study illustrates the importance of usability testing in creating culturally relevant and targeted PtDAs, particularly for web-based tools designed to be used by diverse populations with varying health literacy and acculturation.

Uptake of antiestrogens for chemoprevention remains low, in large part due to patients and providers viewing the risk of side effects as disproportionately higher than the potential benefits of breast cancer risk reduction. The literature supports that inaccurate perception of risk may lead women to overestimate their own breast cancer risk or the risk of side effects for chemoprevention agents. Thus, one of the goals in designing RealRisks and the corresponding usability studies was to improve the accuracy of users’ perception of risk. The “clicking” games, designed to help users explore pre-set probabilities, were developed based on previous research on risk perception but were not adequately describing risk to study participants in the format they were presented. By updating the feedback and explanations for these games, we were able to improve the application for all users.

Another key feature that differentiates RealRisks from other chemoprevention PtDAs is the use of preference elicitation games to engage users in making decisions based on their personal preferences. In the original version of RealRisks, users found these games difficult to use and reported confusion about their purpose. By redesigning the tool as a series of sliding scales, users are now able to adjust their preferences towards various aspects of genetic testing or chemoprevention in order to make more informed decisions. Furthermore, by adding more options to the scales, users are empowered to make decisions based on clinical and non-clinical factors, such as how chemoprevention will impact their family. By including the preference scales in the final action plan, users will be able to review their personal preferences towards chemoprevention with their family and primary care provider (PCP).

In our study, Spanish and English-speaking participants voiced similar concerns, and differences tended to arise based on health literacy and numeracy rather than acculturation. Spanish-speaking participants rated the PtDA as somewhat less useful based upon the SUS. Research on breast cancer decision-making among low-literacy Hispanic women, such as those targeted by RealRisks, has found that this population tends to report poorer decision outcomes when using a PtDA. PtDAs for this population tend to be more effective when they are culturally appropriate and better integrate clinicians and family members into the decision-making process. RealRisks is designed to provide women with an action plan that they can discuss later with their PCP and family members to facilitate decision-making about chemoprevention. Furthermore, RealRisks will be integrated into clinic workflow with the BNAV tool for providers. In an ongoing pilot study with 50 patient-provider dyads and an upcoming randomized controlled trial of RealRisks for high-risk women and BNAV for PCPs, we will assess shared decision-making when both patients and providers are primed prior to the clinical encounter with these decision support tools.

Strengths and limitations

There were several strengths to our usability studies. We achieved a sufficient sample size to identify at least 80% of usability concerns with RealRisks. The recurrence of certain themes in later studies indicated that we reached the saturation point of usability testing. Features that were problematic during the first iteration of tests, such as the preference elicitation scales, were viewed as more usable in the second round. Usability testing was conducted in
both English and Spanish in order to address concerns relevant to the multi-ethnic population that the PtDA was designed for. Furthermore, the population included in the study was diverse and represented a broad range of ages, ethnicities, and acculturation levels. The majority of problems for English-speaking subjects were also found by Spanish-speaking users, indicating that there may be similarities in how both populations view PtDAs, despite differences in acculturation and health literacy.

Limitations in our study were primarily due to the structure of RealRisks and usability testing in general. The web-based format of the PtDA restricted the inclusion of women who were computer illiterate or unable to read a computer screen. Additionally, the think-aloud protocol was difficult for older participants to understand and utilize and necessitated frequent prompting by the study moderator. While think-aloud protocols are standard for usability testing, there is some evidence that using concurrent think-aloud methodology might decrease user engagement during studies. While our sample was diverse, recruitment was limited to women who were able to come in for testing during work hours and thus may not have fully captured usability issues that would impact all populations that use RealRisks. Finally, the time limit placed on each usability study meant that some subjects were unable to complete the entire PtDA in a single sitting. Participants were guided towards relevant sections when time was low, and therefore certain problems may have been missed by users who skipped modules.

Conclusion

In conducting usability studies for the RealRisks decision aid, our team was able to identify several key issues that informed the design of the tool for our target population of low-literacy, multi-ethnic women. Our study is unique in that it incorporated the suggestions and feedback of a very diverse population of potential users and found that concerns were similar between English and Spanish-speaking participants but that differences arose based on health literacy and numeracy. Users reported moderate to high satisfaction with the overall usability of the tool, and future changes will be made to address concerns specific to Hispanic users. Future usability studies of web-based PtDAs should take into account the computer literacy, acculturation, and health literacy of participants when testing in order to develop tools which are accessible to diverse populations.

References


Prioritization and Refinement of Clinical Data Elements within EHR Systems

Sarah A Collins, RN, PhD\textsuperscript{a,b,c}, Emily Gesner, DNP, RN-BC\textsuperscript{a}, Perry L. Mar, PhD\textsuperscript{a,b,c}, Doreen M. Colburn, RN, MSN\textsuperscript{d}, Roberto A. Rocha, MD, PhD\textsuperscript{a,b,c}

\textsuperscript{a}Clinical Informatics, Partners eCare, Partners Healthcare Systems, Boston, MA; \textsuperscript{b}Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; \textsuperscript{c}Harvard Medical School, Boston, MA; \textsuperscript{d}Accenture Health Practice, Atlanta, Georgia

ABSTRACT

Standardization of clinical data element (CDE) definitions is foundational to track, interpret, and analyze patient states, populations, and costs across providers, settings and time – critical activities to achieve the Triple Aim: improving the experience of care, improving the health of populations, and reducing per capita healthcare costs. We defined and implemented two analytical methods to prioritize and refine CDE definitions within electronic health records (EHRs), taking into account resource restrictions to carry out the analysis and configuration changes: 1) analysis of downstream data needs to identify high priority clinical topics, and 2) gap analysis of EHR CDEs when compared to reference models for the same clinical topics. We present use cases for six clinical topics. Pain Assessment and Skin Alteration Assessment were topics with the highest regulatory and non-regulatory downstream data needs and with significant gaps across documentation artifacts in our system, confirming that these topics should be refined first.

Introduction

Standardization of clinical data definitions for use in a learning health system is a critical foundation to achieve the Healthcare Triple Aim: 1) Improving the experience of care, 2) Improving the health of populations, and 3) Reducing per capita costs of health care.\textsuperscript{1} Improving the experience of care requires effective coordination across people, time, and settings, but is effectively enabled only with consistent data to track patient states. Improving the health of populations is dependent on common data definitions to group similar patients across sites and providers, enabling the identification and tracking of patient need and outcome patterns. Reducing per capita costs of health care requires reliable comparisons of clinical data across settings, health professionals, and research databases, particularly considering the need to understand relationships between cost, care complexity, and patient outcomes.

Prospective clinical data collection, even if performed efficiently, is not substantially useful if it is not “reliable” (i.e. collected using inconsistent data definitions), and/or not “computable” (i.e. collected using nonstandard formats and values). Unfortunately, consistent and standards-based data definitions for EHR clinical documentation do not naturally emerge, even within the same clinical information system, without proper clinical governance and technical oversight.\textsuperscript{2}

Background

Definition of standardized data sets for clinical documentation, often operationally known as Minimum Data Sets, have been used to enable effective clinical data analytics, such as nursing care and management across hospitals, and clinical assessment of all residents in Medicare or Medicaid certified nursing homes.\textsuperscript{3–8} The Nursing Minimum Data Set and Nursing Management Minimum Data Sets have been added to LOINC\textsuperscript{9,10}, an internationally adopted standard terminology. Standard terminologies help ensure that discrete clinical data elements (CDEs) can be aggregated and compared across patients, providers, and care delivery sites. Standardized data sets also identify the specific collection of CDEs necessary to represent a given clinical domain or topic. Each collection of CDEs can be simply characterized as a “reference model,” or more appropriately as a “detailed clinical model”.\textsuperscript{11–13} A few publically available sources provide ready-to-use clinical reference models, or sometimes preliminary examples that can inform the development of new models for a given clinical topic.\textsuperscript{14–19}

At our organization we leverage these terminology and model resources to refine existing clinical data reference models, or to develop new models, always seeking a direct application to our commercially available EHR.\textsuperscript{20} We have developed a process to validate reference models based on pre-defined best practices and consensus of interprofessional subject matter experts (SMEs).\textsuperscript{21} The central deliverable of our process is a reference model that can be effectively implemented using our EHR, ensuring consistent data capture and documentation across settings, professions, and purposes.
Detailed and consistent data capture within commercially available EHRs is critical to building a learning health system. While the need for consistent data is widely recognized, the required methods and scope definition, particularly the specific steps for identifying EHR data collection tools (e.g. forms and flowsheets) that need refinement and alignment with available standards, remain unclear to many stakeholders and organizations. Critical initiatives are in process, such as the S&I Framework Structured Data Capture Initiative and HL7 Clinical Information Modeling Initiative (CIMI). However, individual organizations are overwhelmed with the effort required to refine CDE definitions within their EHRs and may ask “where do we start?” or “do we have enough resources to complete the work?”.

To help organizations answer these questions, we have defined practical analytic methods to: 1) identify priorities for the definition of clinical data reference models and, 2) identify and resolve gaps between existing EHR data collection tools and validated reference models. These methods can be applied within and across EHR systems as a systematic and rigorous path to align data definitions within and, eventually, across healthcare organizations.

Methods

We previously published a description of a process for governance and refinement of structured data elements. In that publication we outlined a 10-step approach that: 1) identifies clinical topics, 2) creates draft reference models for clinical topics, 3) identifies downstream data needs for clinical topics, 4) prioritizes clinical topics, 5) validates reference models for clinical topics, 6) perform gap analysis of EHR CDEs compared against reference model, 7) communicates validated reference models across project members, 8) requests revisions to EHR CDEs based on gap analysis, 9) evaluates usage of reference models across project, and 10) monitors for new evidence requiring revisions to reference model.

In this paper we focus on a more detailed description of the metrics and analytical processes involved in the scoring of downstream data needs to prioritize clinical topics and, the process to identify gaps when EHR documentation artifacts are compared against reference models. Figure 1 illustrates how these criteria combine to help prioritize and direct EHR refinement. We view steps 3, 4, and 6 as the most critical steps from an organizational perspective, since these steps provide a rigorous and repeatable method to identify refinement priorities in the context of limited resources. A detailed description of steps 3, 4, and 6 is not included in our prior publication.

Figure 1. Criteria to Prioritize Clinical Topic Refinement

Downstream data needs

The analysis of downstream data needs begins with an environmental scan for the following types of needs related to the identified clinical topic: 1) regulatory, 2) billing, 3) reporting (via automated extracts), 4) clinical decision support (CDS), 5) reporting (via manual chart review), 6) quality initiatives, 7) past data usage statistics, 8) order sets, 9) plans of care, 10) institutional protocols, and 11) literature (published evidence). The environmental scan may require input from SMEs, such as hospital quality and compliance experts. Since the aim of this analysis is to identify high priority topics, an extensive and exhaustive environmental scan is not always necessary. This approach purposefully exploits the underlying assumption that important downstream data needs are readily and efficiently identified during brief interactions with SMEs and from relatively simple online searches. In exploiting that assumption, the usefulness of this method is tied to expedience and analytical rigor, as well as repeatability for comparisons between topics.

Data needs are categorized in two tiers: “direct data needs” (Tier 1) and “indirect data needs” (Tier 2). A direct data need is defined as requiring a process of automated reuse for data filed within a structured CDE, such as electronic Clinical Quality Measures (eCQMs) to Centers for Medicare and Medicaid Services (CMS) programs. An indirect data need is defined as not requiring a process of automated reuse for data filed within a structured CDE. Rather, an indirect data need may be fulfilled using manual chart review processes, such as chart extraction (e.g., quality initiatives), or in the course of clinical care (e.g., reference to protocols). Each identified data need counts as ‘1’ and is applied to the appropriate data need category (see Table 1).

Usage data is also an important variable in our model. During the transition phase while implementing a new EHR at our organization, we used usage statistics from our legacy EHR systems in our model. Six months after our new
EHR was implemented, we switched to usage statistics from our new EHR system. The usage data statistic is a simple rate of use per month and needs to be normalized for comparison with the other data need counts (which in our experience are typically in single digits). Based on an analysis of the counts of usage per month from our legacy systems for our initial set of clinical topics, we determined a cutoff range of 1,000 data points per month for any CDE to indicate high usage. CDEs that had less than 1,000 data points per month received a score of 0.5, topics with greater than or equal to 1,000 data points per month received a score of 1, and CDEs with no usage per month received a zero. The specific cutoff range and additional scores in the instance of particularly high usage (e.g., score of 2 for 2,000-2,999 data points/month, score 3 for > 3,000 data points per month etc.) may be modified per organization based on typical usage rates taking into account number of EHR users and patient encounters.

The downstream data need ‘count’ for each data need category and the weighted values from Table 1 are applied to the formula in Figure 2 below. The key requirement for this formula was that any topic with regulatory data needs should have a higher score than any topic without any regulatory data needs. In order to accomplish that, the first term in the formula is a regulatory term that includes a factor of the signum function applied to the count of the regulatory data sources. In the event that there are no regulatory data needs, this factor, hence the first term, has a value of zero so that the score value is simply equal to the rest of the formula, which is the nonregulatory part of the score. If there are any regulatory data needs, then the first term has a value which is greater than one and less than two. This property of the first term being greater than one is needed in order to exceed the value of the rest of the formula, or the nonregulatory component, which has a value greater than or equal to zero and less than one. This situation guarantees that a topic with even a single regulatory data need has a higher score than a topic with no regulatory data needs, even if such a topic has many nonregulatory data needs. The inverse square root function was chosen for the formula for more even scaling than an exponential or simple inverse function would allow, so that differences among topics could be better pronounced and visually distinguishable. It was also desired that the formula for this score gracefully scale with added data needs and across topics having different numbers of data needs. Using the inverse square root function in this way allows for this in order to confine the score value to a finite range, rather than allowing an unbounded score value as the number of data sources increases.

Using results from this formula we produce a scorecard summarizing Tier 1 data needs that can be easily disseminated among stakeholders. Regulatory data needs are given the highest weight due to their significance for healthcare institutions. The formula results in a weighted score between 0 and 3. Scores in the range of 0-1 indicate there is a non-regulatory data need only; scores in the range of 1-2 indicate there is a regulatory data need; and scores in the range of 2-3 indicate there are regulatory and non-regulatory downstream data needs. These scores are used to rank topics by their identified downstream data needs.

\[
s = \text{sgn} \left( d_1 \right) \left( 2 - \frac{1}{\sqrt{0.1 r_r + 1}} \right) + 1 - \frac{1}{\sqrt{0.1 r_n + 1}}
\]

where

\[
r_r = w_1 d_1 \quad \text{(regulatory raw score)},
\]

\[
r_n = w_0 d_i \quad \text{(nonregulatory raw score)},
\]

\[
d_i \quad \text{is the count for data source } i,
\]

and

\[
w_i \quad \text{is the weight for data source } i.
\]

Figure 2. Formula to calculate weighted Downstream Data Need Score per Clinical Topic
Gap analysis of EHR Clinical Data Elements compared against Reference Model

The development of reference models loses significance if the subsequent implementation of those models within an EHR system becomes too onerous and is deferred for a later date when implementation resources are available. The purposes of quantifying the gap between current state and the reference model are: 1) to estimate the effort required to close the gap for particular documentation artifacts, 2) to rank the documentation artifacts that should be prioritized for refinement based on estimated effort, and 3) to achieve reliability in alignment with the reference model across documentation artifacts.

Table 2. Calibration Metrics for Gap Analysis

<table>
<thead>
<tr>
<th>Metric</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Mismatch</td>
<td>wrong/total = (incorrect + partial/2 + missing + spurious)/(correct + partial+ incorrect + missing + spurious)</td>
</tr>
<tr>
<td>Undergeneration</td>
<td>missing/possible = missing/(correct + partial + incorrect + missing)</td>
</tr>
<tr>
<td>Overgeneration</td>
<td>spurious/actual = spurious/(correct + partial + incorrect + spurious)</td>
</tr>
<tr>
<td>Substitution</td>
<td>(incorrect + partial/2)/(correct + partial + incorrect)</td>
</tr>
</tbody>
</table>

Codes and Definitions:
- CDE = Clinical Data Element
- Match (Correct): CDE on Documentation Artifact = CDE from Reference model
- Partial Match (Partial): CDE on Documentation Artifact ~= CDE from Reference model (variation between definitions exists but are intended to capture same clinical concept)
- Conflicting (Incorrect): CDE on Documentation Artifact ≠ CDE from Reference model
- Extra (Spurious): CDE is on Documentation Artifact but does is not represented in Reference model
- Missing: CDE is not represented on Documentation Artifact but does exist in Reference model

Table 3. Calculations to Estimate Resources for EHR Refinement Based on Calibration Metrics

<table>
<thead>
<tr>
<th>Change Process</th>
<th>Applicable Calibration Metric Categorization</th>
<th>Time Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHR Configuration Build</td>
<td>Not Complex: High undergeneration rate indicating need to add missing clinical data elements (CDEs)</td>
<td>5 min/CDE</td>
</tr>
<tr>
<td></td>
<td>Average Complexity: High substitution or overgeneration indicating need to swap out CDEs or delete superfluous CDEs</td>
<td>25 min/CDE</td>
</tr>
<tr>
<td></td>
<td>Highly Complex: High rate of at least two of three (undergeneration, substitution, overgeneration) indicating need to redesign artifacts</td>
<td>200 min/CDE</td>
</tr>
<tr>
<td>Subject Matter Expert Review (as needed)</td>
<td>Not Complex: no communication or review required</td>
<td>0 hours</td>
</tr>
<tr>
<td></td>
<td>Average Complexity: communication that change occurred</td>
<td>1 - 2.5 hours/DF</td>
</tr>
<tr>
<td></td>
<td>Highly Complex: review of changes to ensure clinical appropriateness</td>
<td>2.5 - 5 hours/DF</td>
</tr>
<tr>
<td>EHR Unit Testing</td>
<td>Not Complex: High undergeneration rate indicating need to add missing CDEs</td>
<td>10 min/CDE</td>
</tr>
<tr>
<td></td>
<td>Average Complexity: High substitution or overgeneration indicating need to swap out CDEs or delete superfluous CDEs</td>
<td>50 min/CDE</td>
</tr>
<tr>
<td></td>
<td>Highly Complex: High rate of at least two of three (undergeneration, substitution, overgeneration) indicating need to redesign artifacts</td>
<td>400 min/CDE</td>
</tr>
</tbody>
</table>
Table 2 outlines our adaption of the MUC-5 (Fifth Message Understanding Conference) Evaluation Metrics, which were originally used to express error rates as part of a scoring system for template instances produced by information extraction systems when compared to manual extraction by humans. Each documentation artifact being evaluated is compared to the reference model by assigning to each CDE one of the following codes: match, partial match, conflicting, extra, or missing. Next, the four metrics are calculated for each documentation artifact: total mismatch, undergeneration, overgeneration, substitution. Each score is on a scale of 0 to 100, with higher scores for greater discrepancies. The total mismatch score is useful for a gross estimate and ranking of the effort required for aligning documentation artifacts with the reference model. The undergeneration score identifies CDEs that are not included on documentation artifacts. The substitution score identifies CDEs from the reference model that are inconsistently implemented in EHR documentation artifacts. This score is useful to identify changes that are slightly more complex than fixing undergeneration scores by simply adding new CDEs, but likely can be completed without further SME input. The overgeneration score identifies CDEs that may be irrelevant to the clinical topic, i.e. truly spurious, and that require additional review by SMEs for possible inclusion.

We view these metrics as useful to compare any reference model (i.e. set of standardized CDEs) to existing clinical documentation artifacts to identify gaps and work toward alignment. These metrics are also useful to directly estimate the level of effort, time, and expertise to align current documentation artifacts to the reference model and can be useful for readiness assessments at sites with limited resources (see Table 3). For example, based on estimates provided by analysts that implement changes to documentation artifacts in our EHR system, we can use the above described results to calculate the total amount of time required to make the necessary changes, including relevant EHR configuration details that might increase the complexity of the change.

Results

Downstream Data Needs

We display our results of the downstream data analyses for 6 clinical topics: 1) Pain assessment, 2) Skin Alteration assessment, 3) Lung assessment, 4) Mental Status assessment, 5) Gait assessment, and 6) Living Situation
assessment (see Figure 3). We also present the detailed ScoreCard for the two highest priority clinical topics: Pain assessment (see Table 4) and Skin Alteration assessment (see Table 5). The clinical topic Pain Assessment resulted in the highest weighted score for downstream data needs with a score of 2.76 out of 3. Skin Alteration Assessment was the second highest weighted score at 2.70 out of 3.0. Lung Exam, Mental Status, Gait, and Living Situation assessments were 2.56, 2.48, 2.42, and 2.2, respectively.

These results demonstrate that all six topics include regulatory requirements, indicating they are all high priority topics, and can be ranked in order of highest priority. The weighted score allowed for continuous updates to the score and ranking of these high priority topics as new needs were identified. For example, our initial list of topics did not include mental status assessment and lung assessment, but as usage data from our EHR became available, these two topics were identified as frequently documented and added to the analysis. This approach allows for an iterative process to identify and analyze downstream data needs with continuous updating of the prioritized list for selection of the next topic to devote resources for refinement.

Table 4. Weighted Score Card for Pain Assessment

<table>
<thead>
<tr>
<th>Reason Data is Needed</th>
<th>High Priority Data Needs</th>
<th>Data Need Tallies for selected Clinical Data Elements**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pain Location</td>
</tr>
<tr>
<td>Regulatory</td>
<td>TJC¹ Standards for assessing and treating pain</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NDNQI² Pediatric Pain Assessment</td>
<td>1</td>
</tr>
<tr>
<td>Billing</td>
<td>Identified by SME³</td>
<td>1</td>
</tr>
<tr>
<td>Quality</td>
<td>QM MU PQRS⁴ Oncology Results Report</td>
<td>1</td>
</tr>
<tr>
<td>CDS</td>
<td>Pain indicated add to Plan of Care</td>
<td>3</td>
</tr>
<tr>
<td>Weighted score per clinical data element</td>
<td>2.23</td>
<td>2.27</td>
</tr>
</tbody>
</table>

*Weighted Score Key: 0-1 non-regulatory data need only, 1-2 regulatory data needs, 2-3 non-regulatory and regulatory data needs. **Table only displays the clinical data elements with weighted score > 1.0; ¹TJC: The Joint Commission; ²NDNQI: National Database of Nursing Quality Indicators; ³SME: Subject Matter Expert; ⁴QM MU PQRS: Quality Measures Meaningful Use Physician Quality Reporting System
### Calculation of Gap Analyses of EHR and EDC Compared Against Reference Model

We present the gap analyses for CDEs from each reference model that were identified by SMEs as important to be present on all clinical documentation artifacts in our EHR. We found that the total mismatch rate ranged from 26% to 82% for Pain Assessment CDEs on documentation artifacts. In comparison, the total mismatch rate ranged from 58% to 98% for Skin Alteration Assessment CDEs. Based on these calibration metrics and resources available a threshold should be selected and refinement efforts targeted at any CDEs above that threshold (higher scores = greater discrepancies). For example, we could establish a threshold of 60% for our data set. The undergeneration rate for the CDE ‘Pain Location Qualifier’ was 64% and the substitution rate was 67% across documentation artifacts analyzed. These rates were based on 20 documentation artifacts in which the CDE ‘Pain Location Qualifier’ was missing and 11 documentation artifacts in which the CDE ‘Pain Location Qualifier’ was a partial match. We can calculate (based on Table 3) that the 20 missing CDEs require a total of 100 minutes of EHR configuration (5 minutes/CDE), 0 hours for SME review, and 200 minutes for unit testing (10 minutes/CDE). The 11 partial match CDEs require 275 minutes (25 minutes/CDE), 1-2.5 hours for SME review, and 550 minutes for unit testing (50 min/CDE). Therefore to optimize the clinical data capture of the CDE ‘Pain Location Qualifier’ throughout our EHR will require a total of about 20 hours of effort. These estimates are intended to be realistic with time allotted for troubleshooting and team communication as needed.

### Table 5. Weighted Score Card for Skin Alteration Assessment

<table>
<thead>
<tr>
<th>Reason Data is Needed</th>
<th>High Priority Data Needs</th>
<th>Data Need Tallies for selected Clinical Data Elements**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Skin Alteration Type</td>
</tr>
<tr>
<td>Regulatory</td>
<td>CMS¹ Pressure Ulcers Acquired After Admission</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>AHRQ² Patient Safety Indicators</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TJC Surgical Care Improvement Project</td>
<td>1</td>
</tr>
<tr>
<td>Billing</td>
<td>Identified by SME</td>
<td>1</td>
</tr>
<tr>
<td>CDS</td>
<td>Pressure Ulcer Present on Admission; Plan of Care Problem</td>
<td>1</td>
</tr>
</tbody>
</table>

*Weighted Score Key: 0-1 non-regulatory data need only, 1-2 regulatory data needs, 2-3 non-regulatory and regulatory data needs. **Table only displays the clinical data elements with weighted score > 1.0; ¹CMS: Center for Medicare and Medicaid Services; ²AHRQ: Agency for Healthcare Research and Quality; ³TJC: The Joint Commission

427
Figure 4. Mismatch of Clinical Data Elements in Documentation Artifacts Compared to Pain Reference Model

Figure 5. Mismatch of Clinical Data Elements in Documentation Artifacts Compared to Skin Reference Model
Discussion

Refinement of EHR CDEs and documentation artifacts is a continuous process that requires consistent effort, yet is often delayed due to a lack of resources, competing deadlines, and difficulty identifying “where to start”. We found that analyzing downstream data needs at the clinical topic level, and calibrating how well documentation artifacts for that topic matched CDEs from a validated reference model was an effective and productive exercise to initiate and direct refinement efforts for our EHR. This approach is both rigorous and flexible. The metrics provide rigorous comparisons while the ranking and selection of clinical topics and CDEs in which to focus our efforts can consider contextual organization factors such as available resources, strategic quality and safety initiatives, and usage statistics. We believe this approach is useful for topic prioritization and re-prioritization as priorities shift. For example, ‘Pain Assessment’ may be shifting in priority and/or management approach due the Opioid Epidemic.

One of the strengths of the downstream data scoring system is that the scoring is performed at the individual CDE level to identify priorities within clinical topics. The clinical topics serve to scope and group the work to catch inconsistencies in CDE definitions within the context of that topic. For example, inconsistencies may not be detected if only the CDEs that address regulatory requirements were selected for analysis and refinement. For example, a Pain Assessment Scale may appropriately capture data to meet a regulatory need; however, if the other CDEs on that documentation artifact are inconsistently defined it will be challenging for an organization to use sets of data from that documentation artifact in a useful way, besides meeting that isolated regulatory requirement.

To our knowledge, this is the first publication to propose a prescribed, rigorous process and metrics to use for prioritizing work to optimize EHR structured CDEs. We believe the initial time investment required to conduct these analytic activities is realistic and is particularly critical for organizations with limited resources. As expected, we found a high rate of mismatch between a number of CDEs from the reference models and their implementation on EHR documentation artifacts. We used these metrics to highlight which CDEs have the greatest need for alignment with EHR documentation artifacts and the resources required to achieve that alignment. As one example, we quantified that 20 hours of resources was required to optimize one CDE throughout the EHR system. These estimates assume that work is conducted to refine only one CDE at a time. In the instances when refinements can be scaled (a set of CDEs can be refined on one or more documentation artifacts together) there is likely a decrease in total resources required. From a governance perspective, the ability to quantify the resources needed to refine a constrained set of high priority clinical topics is important for organizations hoping to realize a return on investment from the interoperability and data analytics capability of their EHR implementation. The priorities and estimates that result from our analytics can be used to prioritize refinement work to fit allotted resource hours or to request resource hours based on refinement work that is a high priority for an organization.

Limitations

This work has been performed at one integrated health care system during, and immediately after, implementation of a vendor-based enterprise-wide EHR. These metrics and processes should be applied at other institutions to understand how to best apply and integrate them across different operational clinical informatics governance structure and resources. Evaluation of the use of these metrics and processes is needed to quantify their added value in the context of enabling analytics and secondary use of structured data as well as prioritizing work and resources. Finally, the work described is limited by the topics identified and the scope of those topic areas. Topic identification could be enhanced through mining or Natural Language Processing of EHR data for topic areas that are highly used.

Conclusion

Refinement of CDE definitions on documentation artifacts across an enterprise EHR system can be an overwhelming task. We propose a set of metrics that can prioritize and direct this work as part of continuous refinement activities and alignment of EHR data. We presented use cases of high priority topics. In the use cases presented, we found high rates of mismatch between CDEs from our validated reference models and documentation artifacts, indicating the need to move forward with refinement of these artifacts. EHR refinement is an iterative process that requires insightful expertise and motivated individuals. These metrics can illuminate priorities and gaps to focus efforts. We believe these analytical steps are practical and generalizable across organizations and should be shared as a critical collaborative step toward continuous EHR refinement and alignment.

Acknowledgements

We would like to acknowledge the Structured Clinical Data Element Workgroup from the Partners eCare project and members of the subject matter expert panels that contributed to this work.
References

Forecasting the Spread of Mosquito-Borne Disease using Publicly Accessible Data: A Case Study in Chikungunya

Kathryn M. Cooper, Ph.D.1, Dhundy R. Bastola, Ph.D.2, Robin Gandhi, Ph.D.3, Dario Ghersi, M.D., Ph.D.4, Steven Hinrichs, M.D.5, Marsha Morien, MSBA 6, Ann Fruhling, Ph.D., MBA7

1-4,7 School of Interdisciplinary Informatics, University of Nebraska at Omaha, Omaha, NE, United States; 
5Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE, United States; 
6Department of Health Services Research and Administration, College of Public Health, University of Nebraska Medical Center, Omaha, NE, United States

Abstract

Mosquito-borne diseases account for multiple public health challenges in our modern world. The international health community has seen a number of mosquito-borne diseases come to the forefront in recent years, including West Nile virus, Chikungunya virus, and currently, Zika virus. Predicting the spread of mosquito-borne disease can aid early decision support for when and how to employ public health interventions within a community; however, accurate and fast predictions, months into the future, are difficult to achieve in urgent scenarios, particularly when little information is known about infection rates. New sources of information including social media have been proposed to accelerate the development of predictive models of disease progression. In this research, we adapted a previously described model for the spread of mosquito-borne disease using open intelligence sources. The novel implementation of a mixed-model for mosquito-borne disease was capable of being executed in minimal runtime. The results indicate that this model yields fast and relevant results with acceptable margins of error.

Introduction

In August 2014, in response to the rising threat of the Chikungunya virus in 47 Pan-American countries, the United States Defense Advanced Research Projects Agency (DARPA) announced its “Forecasting Chikungunya” Challenge, supported by Innocentive.com1. The goal of this challenge was to gather researchers, analysts, and interested parties to investigate and predict how the Chikungunya virus (CHIKV) would impact populations in 55 countries with relatively untested immunities, such as the United States2. Strategies for preventing the spread of Chikungunya on an individual level typically include avoiding travel to emergent areas, wearing clothing that covers extremities and exposed skin, using mosquito nets for beds, and of course, the usage of repellent3. Public health officials can encourage prevention as well by distributing mosquito nets, spraying densely mosquito-populated areas, and while notifying the public at large of methods for prevention. However, with CHIKV threatening the estimated 950 million individuals currently living in Pan-American countries, it is helpful to determine where the virus may hit hardest in tandem with, if not before, implementation of prevention strategies. There are a number of epidemiological strategies for modeling the spread of mosquito-borne diseases such as dengue fever4, malaria5, and yellow fever6. In 2012, Ruiz-Moreno et al. introduced a combined climate and epidemiological model for predicting the spread of Chikungunya virus in the United States via the Aedes mosquito, finding that climate-based changes could be used to classify regions of the United States according to epidemic risk7. In this model, Ruiz-Moreno et al. incorporated parameters for mosquito and human population density, temperature, and initial infection rates to forecast the potential peak of CHIKV with enough accuracy to recommend potential areas for targeted public health interventions, should the need arise.

To address DARPA’s call for methods to forecast the spread of CHIKV, we have implemented and modified Ruiz-Moreno’s model to extend it for use in any country or region, but specifically implemented it for Pan-American countries. The model was implemented for the months of September 2014 to February 2015, as per DARPA’s solicitation. This model takes as input total country population (N), land area (km²), mosquito population (M), temperature (C) and number of infected individuals (I). The population and area variables can easily be computed using available online resources (described further in Methods) and mostly remain constant, so the model largely depends on the current temperature (also easily found online, but frequency of change is according to the user), the mosquito population (estimated) and the number of infected individuals. In the Methods section, we describe how
we are able to collect these parameters using publicly available data and calculate or infer other parameters required to run the model. We have implemented the model using the R programming language to discern the rise in infected individuals, per country, on a weekly basis.

Methods

To forecast the number of cases infected with Chikungunya Virus (CHIKV) across the Americas, we modified the model proposed by Ruiz-Moreno et al\(^7\). The modified model, which we will informally identify as our SEIRM model, uses ordinary differential equations to simulate the dynamics of mosquito-borne infections in human populations. In particular, the human population is divided into susceptible (S), exposed (E), infected (I), and recovered (R) individuals. The mosquito population is also subdivided into immature eggs, larvae, eggs under diapause, and mature susceptible, exposed and infected. Parameters related to the life-cycle of the *Aedes* mosquito were taken from the literature, as reported in Ruiz-Moreno et al\(^7\). All of the data sources described below used for the forecasts are freely available and readily accessible.

Data Sources

The main data sources used included information from news reports gathered via an automated news information system (solely from HealthMap.org), weather data, and population and country size data. The model requires the following initial conditions: 1) the number of susceptible, exposed, infected, and recovered individuals in a given country; 2) the size of the mosquito population. The number of infected cases per country was extracted from HealthMap (www.HealthMap.org), an automated system that monitors information sources on outbreaks\(^8\). The rationale behind choosing HealthMap is that it provides almost instantaneous reports at a local level, as opposed to more traditional surveillance systems that may be lagging behind in their reports. This information can be filtered by species, disease, dates, and location; the data collected particularly for CHIKV was reported as it happened and not on a regular interval (i.e. a country was only reported to have cases on HealthMap if cases were found that day, week, or month). HealthMap was the main source for CHIKV infection numbers, but it does have some disadvantages. In some cases, HealthMap provides case numbers for entire countries and cities within those countries; the numbers provided for major cities might be contained within the country level estimate. To account for this, only numbers counted for entire countries were considered for this project. A second issue is that in some cases numbers were reported based on data mined from news reports, resulting in inaccurate reports on HealthMap. For example, in an article that may have stated “Experts believe cases could reach up to 100,000 in Chile,” HealthMap may report 100,000 cases for Chile when the actual number of cases would be much lower. This is an issue that seems to have been resolved on the website but was present in the early stages of the project. Some countries never had reported numbers of cases, or in the case of the Caribbean Islands, numbers were reported in groups (i.e. “St. Bart’s, St. Vincent, and St. Kitts collectively have X amount of cases.”) Finally, due to inconsistency of CHIKV reporting, there could be a long absence of reported cases for any given country and then a week where new numbers were reported everyday; the question remained which number to be used. As a result of these issues, we adopted the following guidelines for collecting CHIKV infection numbers from HealthMap:

1. For a given month, if a country reported infected cases on CHIKV, the *most recent* number of cases was recorded for our forecast (even if this number was not the highest possible).
2. If there was a large deviation from previous numbers (for example, one case reported in August and then 50,000 cases reported in September), the articles were manually investigated for their veracity – the *most recent* number with the most likely cases reported was recorded for use in the forecast.
3. Only numbers reported for countries were used; numbers for individual cities contained within eligible Pan-American Health Organization (PAHO) countries were ignored.
4. If a country had no cases reported for that time interval, a Google News search for the country’s name and chikungunya (i.e. “Anguilla chikungunya”) was performed to validate that no cases had been reported. If an article was found where some cases had been reported, it was individually evaluated for reliability and if it was deemed a reputable source, that number was recorded. If this scenario occurred, the source of the number recorded was also recorded.
5. If no change had been reported from month to month for a country via HealthMap and no reputable new source was found via a Google News search, the initial number of infected was not changed.
6. For each country at each forecast, we have recorded whether the number of infected cases came from HealthMap or a news article found via Google News search.

The CHIKV numbers were recorded as described above via HealthMap and other sources were used for each country as the initial number of *Infected*. To estimate the other initial values (*Susceptible, Exposed, Recovered*), the
total population was found for each country via Google search (i.e. “Anguilla population”) and recorded using the following approach. Given a list of $C = \{c_1, c_2, \ldots, c_N\}$ countries provided, a Google search for each country in $C$ was performed using the search term “$c_i$ population” where $c_i$ represents the current country in $C$. This search was done manually as less than 100 countries were being studied and to use an API would require more time and work than just searching manually. For country population, Google provides not only a list of search results in the form of hyperlinks, but also a “definitive” population based off the last major formal census of the region as the top result. This is the number that was used to determine the total country population – the definitive population stored by Google’s “Map Data” resource. The values for remaining compartmental populations of a country (Susceptible, Exposed, and Recovered) were estimated using the following simple equations (chosen arbitrarily to reflect reasonable and conservative estimates):

Given the Country population ($P$) and Infected Population ($I$):

$$\text{Recovered (R)} = I \times 0.75$$
$$\text{Susceptible (S)} = (P - I - R) \times 0.9$$
$$\text{Exposed (E)} = (P - I - R) \times 0.01$$

Initially, the Exposed population was set to $(P-I-R) \times 0.1$ to ensure that $S+E+I+R = P$. However, in our model this resulted in a very large number of persons becoming infected and was felt to inaccurately describe a realistic scenario of exposed persons. As such, in October forecasts and further, a new Exposed equation was adopted and each country was held to the condition that $S+E+I+R < P$.

To estimate the size of the mosquito populations at a per-country level, we used mosquito density data reported by Eiras and Resende for Brazilian municipalities\(^9\), and the surface area of each country. Although the estimates are necessarily only approximations, we found a change in predictive power over simply using a fixed number of mosquitoes for each country.

The average monthly temperature (AMT) in Celsius was collected for each country for every month from 2001 to 2010. Latitude, longitude, and temperature data was downloaded from the following website provided by The Centre of Climatic Research, University of Delaware\(^10\). The Google maps API was used to convert the latitude and longitude information to its respective country name. Temperature data for a country was calculated as the average of temperature readings from all the points in that country. For each month from 2001 to 2010, the AMT was averaged to get a single value for input into the model. As such, for every forecast, the input parameters for weather changed according to the average of all AMTs for that country in that month from 2001 to 2010.

**Model Implementation:**

The aforementioned modified Ruiz-Moreno model was implemented in R using the `deSolve()` library and an array of inputs and parameters described below. Equations that have been modified from the Ruiz-Moreno manuscript are highlighted in bold and described in a comment; many of these modifications are very small. All involve removing a density-dependent parameter specified in the original paper that was defined only for certain US cities (Miami, Atlanta, and New York) and these parameters could not be effectively estimated for duplication in our model. Very brief pseudocode of the model can be found below. The code for the model can be found at https://github.com/katecooperOMA/Chikungunya.

```r
/***Begin model***/
1 Load library deSolve and other requirements
2 /***Load inputs by user***/
3 initials = {
4   country = user input
5   S = user input #Susceptible
6   E = user input #Exposed
7   I = user input #Infected
8   R = user input #Recovered
9   G = user input #Mosquito population by country size
10  temp = user input #Average AMT for country during current month
11 }
12 /***Load Hard Parameters by Ruiz-Moreno***/
13 parameters = { Adult mosquito population, symptomatic/asymptomatic ratio, etc.  }
14
15 /***Load Soft parameters by Ruiz-Moreno #See [7]for full definitions***/
16 parameters = { mosquito egg mortality, larval mortality, incubation period, etc.}
17 population = S+E+I+R
```
19 \( dT = \text{sequence 1 to 4 by 1/7} \)  \# define the time interval desired
20 \text{SEIRM}\_function = |
21 \quad \text{define changes in mosquito populations (adult, larval), compartmental populations}
22 \quad \text{population} = S+E+I+R
23 \quad \text{der}<-\text{c}(dS, dE, dI, dR, dG, dD, dL, dAs, dAe, dAi) \quad \# \text{der is a function of deSolve library}
24 \}
25
26 \text{simulation}<-\text{as.data.frame}(\text{lsovda}(\text{initials}, dT, \text{SEIRM}\_\text{function}, \text{parameters}))
27 \text{output results}

For each forecast, the inputs (S, E, I, R, G, country name, and temperature) were collected and ran using a small Perl wrapper that iteratively ran each command in a designated folder where outputs and error logs, if any, were collected. The model gives S, E, I, and R predictions for 21 weeks after the initial values but only the Infected numbers are recorded. Because the country temperature input changed from month to month, the model was run for every month remaining in the forecast. For example, for the December 1 forecast, the model was run three times: once for December, once for January, and once for February. The output of December’s forecast, then, became the input for January’s forecast, i.e., if the number of infected cases for the last week of December was predicted to be 30, then 30 was used as the initial value of Infected for the January iteration of the December forecast. Similarly, the last week of the January forecast then became the initial value for the February model iteration. This was structured according to the DARPA Forecasting Chikungunya Challenge guidelines.

A number of parameters used in our model were held steady at every iteration; these included: the adult mosquito population, the number of eggs per mosquito per day, the rate of symptomatic to asymptomatic individuals, the biting rate of the mosquito, the infective period, the probability of both human-to-mosquito and mosquito-to-human transmission, the density dependent factor for reproduction and density function, and the natural and disease-induced mortality rates. The numbers are estimates and were taken from Ruiz-Moreno et al. and after discussion, it was determined that none of these parameters had enough evidence to warrant their change in our model in its early stages. Other constant parameters included were country population and area/size.

Saint Martin and Sint Maarten were counted as two separate countries.

Results

Real-world Performance. The ability to effectively display the results of simulations is particularly important in this domain, where several different scenarios might have to be explored and interpreted by professionals with diverse backgrounds and expertise. Our model generates an easily interpretable time-dependent chart, illustrating the number of predicted susceptible, infected, and recovered cases over a given time range.

In Figure 1, we documented our forecasts for all months (September, October, November, December, January, and February) compared to Historical PAHO data in Haiti. For example, our September forecast for Haiti predicted a slow rise to around ~30,000 infected cases reported by February (Figure 1); by the time the October forecast came due, HealthMap had reported several cases in Haiti and future predictions reflected this update. With a population

![Figure 1. Our forecasts for Haiti compared to Historical PAHO data (HIST)](image-url)
around 103 million, the number of cases in Haiti was predicted to represent around 90,000 or 0.87% of the population. This prediction model also applies well to smaller countries. A similar data plot is shown below for Montserrat (Figure 2), which has a population of around 5,900 individuals. Our initial predictions for September and October predicted a rise to approximately 100 or 120 infected individuals by February; in the final week of December, Montserrat reported an increase in cases from 6 (first reported in Week 42) to 119 (reported in Week 52). However, with few to no cases reported by HealthMap in December or January and a change in temperature, our December, January, and February forecasts assumed a low initial Infected value and a slow climb in the number of cases. Still, at 119 cases, this represents about 2% of Montserrat’s total population.

St. Bart’s (Figure 3) is an example of a case where nothing was being reported specifically for the country in the news or on HealthMap itself; occasionally it would be included in a count for a group of countries, but being a small island with a small population, this country did not manifest itself in the news often. As a result, our forecasts predicted a slow rise in the number of cases (compared to the actual number of cases reported); when these predictions did not reflect the HealthMap/news data released in the next month, the slow rise in cases was simply pushed back by the model to the next month.

For countries where the mosquito population is not regularly be exposed to CHIKV, such as the United States or Canada, our initial forecasts predicted far too many cases because numbers for exposure and mosquito populations were not modified (Figure 4). However, once the temperatures in these countries dropped, our predictions became (slightly) more accurate.

Suriname (shown in Figure 5) is an example of a country where we are not yet quite sure how well our predictions have come out; as of the time of this manuscript they had not reported their CHIKV cases since Week 44. All indicators suggest that while we had a few differences in the severity of infectiousness trend, this prediction may be one that is more accurate than others. Further, we can also potentially display the number of cases on a geographic map over time, a type of representation that is particularly useful to track the spread of the disease and the effects of containment measures in specific areas.

Sensitivity Analysis. It is difficult to perform accurate sensitivity or uncertainty analysis on our model due to incomplete and inconsistent PAHO reporting. However, we have completed some tests to investigate how much our forecast would deviate by modifying temperature, mosquito population, and number of infected for the most populous country (United States at 313 million), the least populous country (Saba at 1,824), the median country for population (Guadeloupe at 17 million) and the average country for population (Chile at 405,739). We ran our forecast for 22 weeks based on our input data for February 2015 for each of these countries, and identified three major areas of modification to investigate: (1) change in the original input temperature (+ or -20C), (2) change in mosquito population (+ or – 50% of the February input mosquito population), and (3) change in the number of infected (increases of 10%, 20%, and 30% of the February input of infected cases). The results of these iterations are below.
Saba: the smallest PAHO country by population (Figure 6). The number of infected for Saba was only distinctly affected by the temperature – it would appear based on the curves for both an 20C increase and decrease in temperature that the ideal temperature for CHIKV spread would be around the current average temperature there, 27.2C. However, the forecast shown assumes this temperature remains the same for the next 20 weeks – this would not be the case in a real world situation. Changes in mosquito population and number of infected have minimal effect on the outcome of the 22 week forecast, with mosquito population having a 22 week range of ~25–45 infected and number of infected forecasts all having a 22-week prediction of ~35 infected.

Figure 3. Our forecasts for St. Bart’s compared to Historical PAHO data (HIST)

Figure 4. (Left) All forecasts for the United States and (Right) the December, January, and February forecasts for the United States

Figure 5. Our forecasts for Suriname compared to Historical PAHO data (HIST)
Chile: the average PAHO country by population (Figure 7). None of the forecasts for Chile were distinctly affected by any change in parameter. A change in temperature resulted in a 22-week prediction range of 5966 infected (20°C decrease) to 7158 for a 20°C increase, a difference of only 1192. A change in mosquito population resulted in a range

Figure 6. Modified Forecasts for Saba. Top left: Forecasts for original Feb. 2015, or with a + or -20°C input modification. Top right: Forecasts for original Feb. 2015, or with a + or -50% mosquito population. Bottom: Forecasts for original Feb. 2015, or with a + 10%, 20%, or 30% population increase.

Figure 7. Modified Forecasts for Chile. Top left: Forecasts for original Feb. 2015, or with a + or -20°C input modification. Top right: Forecasts for original Feb. 2015, or with a + or -50% mosquito population. Bottom: Forecasts for original Feb. 2015, or with a + 10%, 20%, or 30% population increase.

Chile: the average PAHO country by population (Figure 7). None of the forecasts for Chile were distinctly affected by any change in parameter. A change in temperature resulted in a 22-week prediction range of 5966 infected (20°C decrease) to 7158 for a 20°C increase, a difference of only 1192. A change in mosquito population resulted in a range
of 6222 to 7990 infected for a 50% decrease and a 50% increase respectively. The increase in number of infected resulted in a predicted range of 7208 (10% increase) to 7360 infected (30% increase).

Guadalupe: the median PAHO country by population (Figure 8). None of the forecasts for Guadaloupe were
distinctly affected by any change in parameter; in fact, two of the modifications (temperature and mosquito population change) had almost no impact on predicted number of infected at all. A change in temperature resulted in a 22-week prediction range of 1,101 infected (20°C decrease) to 1,093 for a 20°C increase, a difference of only 8. A change in mosquito population resulted in a range of 1,097 to 1,112 infected for a 50% decrease and a 50% increase respectively. The increase in number of infected resulted in a predicted range of 1,203 (10% increase) to 1,402 infected (30% increase).

United States: the largest PAHO country by population (Figure 9). The forecasts for the United States were the most affected by any changes both temperature and mosquito population. A change in temperature resulted in a 22-week prediction range of 9,333 infected (20°C decrease) to double that at 18,556 for a 20°C increase. However, the original February forecast inputs were very similar in range to the 20°C decrease in temperature, suggesting that the 20°C increase or decrease signaled a change in mortality for mosquito populations, thus affecting the number of infected. A change in mosquito population resulted in a range of 5,992 to 11,048 infected for a 50% decrease and a 50% increase respectively. The increase in number of infected resulted in a predicted range of 8,574 (10% increase) to 8,470 infected (30% increase).

These results suggest that our model is potentially at its best when dealing with populations in the average or median range, but can also perform on very small and very large populations. The biggest changes observed were between the parameter modifications in the United States, the largest PAHO country by population. Compared to the forecasts of Guadaloupe (with a population of 17 million), this suggests that a threshold for integrity of the model as it sits currently may lie somewhere between 17 and 300 million population. All of this, of course, is speculative at best – more rigorous robustness testing with accurate PAHO data would better reveal the strength and predictive power of the model.

Applicability. The SEIR model that was used for our predictions can be easily applied to other mosquito-borne diseases such as Malaria, West Nile, dengue fever, Congo-Crimean hemorrhagic fever, or yellow fever. Depending on the type of disease, parameters related to the infectivity of the agent and to the life cycle of the vector would have to be suitably chosen, but the general structure of the model does not require major changes. This extends to diseases transmitted via other vectors such as ticks to humans and animals alike; the main dependencies of the model lie in initial values, population, temperature, and area of a country or district. Of course, the model can be modified. In fact, related SIR (Susceptible, Infected, Recovered) models have been successfully used in the past to model the spread of dengue fever and other borne infectious diseases [4][5]. This could, then, theoretically be applied to other diseases such as Ebola or measles, with relatively small changes.

Computational Resources. The model does not require specialized hardware, and can run on a desktop computer or laptop. It is implemented in the widely used R language, which is freely available for Windows, Mac OS X, and Linux platforms. A simulation takes approximately 15-30 seconds for all countries on a Late 2013 MacBook Pro with 2.4 GHz Intel processor running OS X 10.9.3.

The model is general in its implementation, and can be easily adapted to accept data sources that are different from the ones that were used for this research. For example, instead of news reports from HealthMap the model could use surveillance data or other sources containing the incidence of new cases. Furthermore, all of the data collection and processing can be scripted and automatically run at certain intervals without user interaction.

Conclusion

Our modified Ruiz-Moreno model is a simple and easily interpretable approach that uses freely available data. The datasets used as input are reliable, ranging from very high confidence (population and country area/size) to moderately reliable (HealthMap data and news reports). The model runs very quickly (under 1 minute) on most laptop computers equipped with the R statistical suite and the essential libraries discussed previously. Additionally, our model is discriminatory – it remains free of extraneous variables and data by depends largely on initial values, temperature, population and land area to determine rates of infection. More variables can be added as they are deemed important by public health authorities. The contribution of this research is a novel implementation of a mixed model for -borne disease that executes in minimal (<5 minute) runtime using only publicly available information.

References


11. The DARPA Innocentive Chikungunya Challenge guidelines were only available to Challenge solvers during the tenure of the challenge itself and are no longer available on the website. While we do not have permission to publicly post these guidelines, a digital copy of the guidelines can be shared upon request.
ODaCCI: Ontology-guided Data Curation for Multisite Clinical Research Data Integration in the NINDS Center for SUDEP Research

Licong Cui, PhD\textsuperscript{1,2,3}, Yan Huang\textsuperscript{1,2,3}, Shiqiang Tao, PhD\textsuperscript{1,3}
Samden D Lhatoo, MD, FRCP\textsuperscript{3,4}, Guo-Qiang Zhang, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Institute of Biomedical Informatics, University of Kentucky, Lexington, KY
\textsuperscript{2}Department of Computer Science, University of Kentucky, Lexington, KY
\textsuperscript{3}Center for SUDEP Research (NINDS-funded Center Without Walls for Collaborative Research in the Epilepsies), Cleveland, OH
\textsuperscript{4}Department of Neurology, Case Western Reserve University, Cleveland, OH

Abstract

Sudden Unexpected Death in Epilepsy (SUDEP) is the leading mode of epilepsy-related death. The Center for SUDEP Research (CSR) is an NINDS-funded Center Without Wall’s initiative aimed at prospectively creating a comprehensive clinical research resource for SUDEP. This resource consists of a growing set of data and biological samples of a statistically significant cohort of patients at an elevated risk, best represented by the Epilepsy Monitoring Unit (EMU) patient population. The Informatics and Data Analytics Core (IDAC) of CSR has developed a state-of-the-art informatics infrastructure, to integrate patient data captured in multiple EMU’s at a greatly accelerated pace. Data quality assurance is a priority of IDAC. This paper reports our approach, Ontology-guided Data Curation for Multisite Clinical Research Data Integration (ODaCCI), to address the challenging task of centralized data curation while new data is continuously generated and integrated from distributed sites. ODaCCI leverages the Epilepsy and Seizure Ontology not only for upstream data capture, but also for supporting a range of quality assurance tasks such as data quality monitoring, data update, and data reports. Between October 2014 and February 2016, ODaCCI has integrated phenotypic and electroencephalogram signal data of 629 patients from 7 clinical sites, while supporting continuous and asynchronous data quality enhancement over time.

1 Introduction

Data errors in research databases are prevalent. They include inaccurate interpretation of data in the initial documents and incorrect data entry into databases [1]. Data quality (DQ) in clinical research is a fundamental topic because it impacts the subsequent decisions and conclusions [2, 3]. Researchers have shown that electronic health records (EHR) often contain errors that may affect research results [4]. However, the process of data quality auditing and improvement is challenging because of multimodality, the lack of provenance information, the lack of coordinated versioning and updates, and the time and effort needed.

Data quality assurance is a unique, acute challenge for prospective, multi-site studies such as the National Institute for Neurological Disorders and Stroke (NINDS) funded Center Without Wall’s initiative called the Center for SUDEP Research (CSR [5]). The CSR is a collaboration of 14 institutions, bringing together scientists and physicians to investigate and identify the molecular and structural brain abnormalities underlying SUDEP. The potential discoveries could then be utilized to identify features that could predict and identify patients at risk of SUDEP. Clinical phenotype, electroencephalogram (EEG), imaging and other data are prospectively collected from CSR patients in multiple participating Epilepsy Monitoring Units (EMUs). As the EMUs capturing multi-modal patient data, such data are de-identified and integrated into the CSR central data repository. Data correction and update at the source is less feasible than in the central data repository, sometimes because of the lack of resources, other times because of the lack of specific knowledge about a domain at an individual site. The Informatics and Data Analytics Core (IDAC) of CSR adopted a strategy to perform centralized data curation, by reviewing and cross-referencing linked physiological data such as video-EEGs to enhance data quality.

The paper introduces an ontology-guided approach to support the IDAC data curation strategy, called ODaCCI, by integrating curation and reporting interfaces into the CSR data integration pipeline. A collection of 95 common data elements were identified by CSR domain experts as the main curation targets. ODaCCI consists of an epilepsy domain ontology-guided, web-based, data curation interface to support data auditing, error correction, and data entry, with an interactive reporting interface for auditing the completeness of the types of data uploaded from multiple sites. A total of 629 patients’ data sets have been integrated from 7 clinical sites from October 2014 to February 2016. Among them, 393 are from University Hospitals of Cleveland (UH), 65 from New York University (NYU), 11 from
University of California, Los Angeles (UCLA), 54 from Northwestern University (NW), 44 from Thomas Jefferson University (TJU), 56 from University College London (UCL), and 6 from University of Iowa (UIowa). ODaCCI supports continuous and asynchronous quality improvement while data has been ingested and integrated from multiple clinical sites.

2 Background

2.1 Center for SUDEP Research (CSR)

Epilepsy is the most common serious neurological disorder, affecting 65 million people worldwide [6]. Sudden Unexpected Death in Epilepsy (SUDEP) is one of the leading modes of epilepsy-related death. More than 1 out of 1,000 epilepsy patients die from SUDEP each year [7, 8]. However, the mechanism causing SUDEP is not well understood [5]. CSR is an ongoing NINDS-funded Center Without Walls initiative for Collaborative Research in the Epilepsies [9], to accelerate the understanding of SUDEP by bringing together extensive expertise from 14 institutions across United States and Europe. Due to the low annual incidence of SUDEP (∼1%), cross-institution data sharing are required to collect SUDEP/near-SUDEP data of sufficient statistical significance. This is the mission of CSR, with the goal of recruiting at least 2,500 cases from EMUs in the participating clinical sites to support prospective recruitment and identify possible risks for SUDEP.

2.2 CSR Data Integration Challenges

Major challenges to the CSR data integration include, but are not limited to, the following:

- **Data heterogeneity.** Phenotypic data were captured in disparate formats (e.g., EHR, PDF documents) in different clinical sites. The integration of heterogeneous data is the most critical issue in a multisite research setting.

- **Data access restriction.** This involves the protection of privacy for patients data. Researchers from one site should not have access to the protected health information (PHI) of the patients from other sites (e.g., patient name, date of birth) without additional IRB review and agreement.

- **Multimodal data linkage.** Phenotypic data and electrophysiological signal data recorded in different devices need to be properly linked.

- **Data quality.** Quality data is essential for use in clinical research [10]. However, in a multisite research effort, inconsistent codings are often seen across different clinical sites. It is unavoidable that unintended erroneous input may happen during the manual data entry process. Missing data is another major issue.

2.3 Related Efforts to CSR Data Integration

To minimize the potential format heterogeneity, a uniform electronic data capturing system, called OPIC (Ontology-driven Patient Information Capture system) [13], has been adapted and enhanced to prospectively capture patient phenotypic data for CSR. To avoid inconsistent coding, an enhanced version of OPIC has incorporated standardized coding for terms in the Epilepsy and Seizure Ontology (EpSO) [12]. In addition, automatic data validity checks (e.g., data must conform to an expected format) have been implemented to ensure better data quality at the data entry stage. The enhanced OPIC has been deployed as a Virtual Machine (VM) image in the clinical sites. These OPIC instances operate behind hospital firewalls.

To address the data access challenge, a central CSR data repository has been built to integrate and store de-identified patient data from multiple sites. A data de-identification module has been implemented in OPIC to remove PHI information before integrating into the central repository. To ensure patient-level record linkage without revealing PHI, a distributed study identifier generation algorithm has been implemented in OPIC using randomized n-gram hashing [11]. At the moment patient record is initially created, a unique study identifier is automatically generated to facilitate the data de-identification and linkage of multi-modal data.

2.4 CSR Data Quality Assurance

Although certain data quality assurance measures have been incorporated in the enhanced OPIC (e.g., data validity check, standardized coding), inconsistencies may still exist due to unintended manual input or misinterpretation of observations from electrophysiological signals. Review, auditing, and curation of the acquired data in the central
repository by CSR domain experts are necessary for ensuring data quality. Inconsistencies found by observing and analyzing electrophysiological signal data will also be used to help curate phenotypic data. Moreover, since patient data is prospectively collected and continuously growing, the central data repository needs to incorporate both incoming data from each individual site as well as curated data by CSR domain experts. Therefore, there is a unique need of a streamlined data integration and curation workflow to facilitate the CSR data integration and quality assurance framework. This paper introduces ODaCCI to address this need.

3 Methods

Figure 1: CSR phenotypic data integration and curation workflow.

3.1 Data Integration and Curation Workflow

Figure 1 shows the overall architecture of ODaCCI, for integrating and curating patient phenotypic data from an individual site to the CSR central repository. Initially, the patient phenotype data is entered into the OPIC system with identifiable information at each individual site. Such phenotypic data includes patient demographics, history, medications, diagnosis, epileptogenic zone, seizure semiology, and EEG findings. Then, automatic de-identification by OPIC and manual de-identification by study personnel are performed to produce de-identified data, which is transferred to the CSR central repository through SSH File Transfer Protocol (SFTP) on a weekly basis. The de-identified data for a single clinical site are then imported to a standalone MySQL database in the central repository, which is interfaced with the web-based data curation system for a domain expert to perform auditing and curation. The curated data is saved in a separate MySQL database that only contains curated data. The rationale for separating the curated data from the original integrated data is to accommodate both the continuously acquired data from the source clinical site and the existing curated data in the central CSR repository. Moreover, when integrating new data into the central de-identified database, the curated data in the separate database will replace certain original data that are not curated from the source.

3.2 ODaCCI: Ontology-guided Web-based Data Curation System

In this subsection, we focus on introducing each of the following components in ODaCCI:

- **Common data elements (CDEs)**, which are data elements that are common to all the individual clinical sites.
- **Ontology-based vocabulary**, which serves as the main knowledge source for possible values of domain-specific data elements, and plays an important role in dynamic generation of web-based curation widget and MySQL statement to interacting with the backend database.
- **CDE to data source mappings**, which link common data elements to actual tables and columns in the backend databases from different sources. Such mappings not only drive the rendering of the backend data in the web
interface for curation, but also facilitate the translation of content captured in the web interface to MySQL statements to be executed in the backend database.

- **Dynamic generation of data curation widget**, which provides a systematic way to generate web-based curation widgets for different types of data elements.
- **Dynamic generation of MySQL statements**, which automatically translates the curation result obtained from the web-based widgets to backend MySQL statements.
- **Data auditing measures**, which provide intuitive and interactive web-based reporting interfaces for domain experts to perform data auditing and curation.

### 3.2.1 Common Data Elements (CDEs)

A set of common data elements from all the data sources were selected by epilepsy domain experts for data integration. Selected epilepsy patient phenotypes include age, gender, epileptogenic zone, etiology, semiology, epileptiform discharge, drug, transection, body mass index, sleep position, smoking, and bleeding. Table 1 shows a partial list of common data elements organized in sections. For example, **DEMOGRAPHIC** section includes data elements **Age** and **Gender**; and **CLASSIFICATION OF PAROXYSMAL EPISODES** section consists of **Epileptogenic zone**, **Etiology**, **Semiology**, **Nonepileptic semiology**, and **Lateralizing sign**.

<table>
<thead>
<tr>
<th>Section</th>
<th>Common Data Element</th>
<th>Name</th>
<th>Type</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEMOGRAPHIC</td>
<td>Age</td>
<td>numerical</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td>categorical</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>CLASSIFICATION OF PAROXYSMAL EPISODES</td>
<td>Epileptogenic zone</td>
<td>ontological</td>
<td>modifier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Etiology</td>
<td>ontological</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semiology</td>
<td>ontological</td>
<td>modifier, only during admission</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonepileptic semiology</td>
<td>ontological</td>
<td>modifier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lateralizing sign</td>
<td>ontological</td>
<td>modifier, only during admission</td>
<td></td>
</tr>
<tr>
<td>EVALUATION</td>
<td>Epileptiform discharge</td>
<td>ontological</td>
<td>modifier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonepileptiform abnormality</td>
<td>ontological</td>
<td>modifier, only during admission</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MR/CT status</td>
<td>categorical</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EEG type</td>
<td>categorical</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>PAST AND CURRENT MEDICATIONS</td>
<td>Drug</td>
<td>ontological</td>
<td>time taken</td>
<td></td>
</tr>
<tr>
<td>EPILEPSY SURGERY FORM</td>
<td>Corpus Callosotomy</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transection</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deep brain stimulation</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>CHECKLIST</td>
<td>Body mass index (BMI)</td>
<td>numerical</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sleep position</td>
<td>categorical</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cardiac disease</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>FOLLOW-UP FORM</td>
<td>Still having seizures</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smoking</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drinking alcohol</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>SUDEP FORM</td>
<td>Frothing around mouth</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fallen out of bed</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bleeding</td>
<td>boolean</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>

The value set (i.e., set of possible values or responses) of a common data element determines the type of the common data element. Types of common data elements include **boolean**, **categorical**, **numerical**, and **ontological**. A data element is called **ontological** if its value set originates from a vocabulary of ontological terms organized as a hierarchy. For both categorical and ontological data elements, their possible values are coded using integers to support effective data capture and retrieval. For instance,

- **boolean**: **Transection** is boolean with value set \{0, 1\} (0 means no and 1 means yes);
- **categorical**: **Gender** is categorical with value set \{1, 2\} (1 means male and 2 means female);
- **numerical**: **Body mass index** is numerical;
- **ontological**: **Semiology** is ontological with value set \{1, 2, \ldots, 51\} (1 means Aura, 2 means Autonomic Seizure, and 51 means Hypnopompic Seizure). The hierarchical view of the value set can be found in Figure 2 (middle).
A common data element may have properties such as modifier (see the last column in Table 1). For example, *Semiology* has two properties: modifier and only during admission. Semiology modifier may be Generalized, Bilateral asymmetric, or Left or Right Axial/Proximal/Distal/Head/Face/Arm/Hand/Leg/Foot. The value of only during admission may be Yes or No. These common data elements as well as their specifications form a common data dictionary managed in a Common Separated Values (CSV) file.

### 3.2.2 Ontology-based Vocabulary

We leverage the Epilepsy and Seizure Ontology (EpSO) [12] as the vocabulary to construct value sets for ontological data elements. For each ontological data element, its value set consists of all the direct and indirect subtypes of a class in the EpSO. For instance, in EpSO, *Epileptogenic Zone* (left in Figure 2) has seven subtypes: *Anterior Head Regions, Generalized, Hemisphere, Multi Focal, Posterior Head Regions, Unknown, and Unlocalizable*. And *Hemisphere* further has subtypes including *Central, Cingulate, Frontal, and Occipital*. For such ontological terms in EpSO, the data capturing system OPIC [13] has dedicated integer codes. Therefore, we use the same standardized codes for data curation in order to seamlessly integrate continuously collected data and curated data.

![Ontology-based vocabulary](image)

Figure 2: Ontology-based vocabulary for *Epileptogenic Zone, Semiology, and Drug*.

### 3.2.3 CDE to Data Source Mappings

Since the data capturing system OPIC has been distributively deployed and continuously operating at each clinical site, disparate sites may have different versions of OPIC, thus the data schemas in the de-identified data sources may not always be the same. Therefore, for each site, the common data elements in the common data dictionary are mapped to source data tables and columns if applicable. For example, for each data source containing patient phenotype *Epileptogenic Zone*, the common data element *Epileptogenic Zone* is mapped to the column “epileptogenic_zone_id” in the table “classification_zones,” and its property *Modifier* is mapped to the column “modifier” in the same table. Such mapping information is maintained in CSV files for each data source.

### 3.2.4 Dynamic Generation of Data Curation Widget

To facilitate manual curation of patient phenotypic data by domain experts, we developed a web-based data curation interface driven by the common data elements and their mappings to data sources. When a user triggers an event to edit or curate a common data element, an interactive dialogue (see Figure 3) is dynamically generated based on the type and properties of the data element as follows.

- If the type is boolean, then a dropdown option of *Yes* and *No* is provided for selection.
- If the type is categorical, then a dropdown list containing the value set of the data element is provided.
- If the type is numerical, then a text box is displayed for editing.
If the type is ontological, then a multi-level dropdown list is rendered according to the ontology-based vocabulary corresponding to the data element. For instance, clicking the “Add Semiology” button inside the interactive dialogue (arrow (a) in Figure 3) first triggers the display of all the direct subtypes of Semiology; clicking one of these subtypes “Aura” further renders its direct subtypes (arrow (b) in Figure 3); clicking the subtype Autonomic Aura further drills down to its direct subtypes (arrow (c) in Figure 3). In addition, the Modifier property of Semiology can be updated by clicking the icon after its value (arrow (d) in Figure 3).

Moreover, the default values for the data elements in the interactive dialogues are automatically populated from the data source. Therefore, the values that do not need curation are kept intact.

The web-based curation interface was implemented using an agile web development environment called Ruby on Rails (RoR) [18] with a MySQL backend database. The integrated and curated data sets are also stored in MySQL databases, which are separate from the main RoR application database. The specifications for the common data elements and ontology-based vocabularies in CSV files are imported into the application database to drive the web interface.

### 3.2.5 Dynamic Generation of MySQL Statement

The common data elements and their mappings to data sources not only drive the rendering of web-based curation interface, but also play an important role in the dynamic generation of MySQL statements for saving the interactive edits provided by domain experts. For each type of common data elements, a general template of MySQL statement is predefined and used for generating the actual MySQL statement for data curation. For example, the general template for an ontological common data element is predefined as:

\[
\text{INSERT INTO } \text{<mapping.table>} \left( \text{id}, \text{<mapping.column>}, \text{<property_1>}, \ldots, \text{<property_n>} \right) \\
\text{VALUES } \left( \text{id_value}, \text{<cde_value>}, \text{<property_1_value>}, \ldots, \text{<property_n_value>} \right);
\]

where <mapping.table> and <mapping.column> represent the data source table and column to which the cde is mapped, and <property_i> represents the i-th property of the cde. All the variables in the angle brackets can be replaced by real values to generate the actual MySQL statement. For instance, in Figure 3, the first record for the ontological common data element Semiology has the following values for the variables in the template:

- <mapping.table>: seizure_type_semiologies
- <id_value>: ‘TSXP606170783305’
- <property_1>: modifier
- <property_1_value>: ‘Left Arm’
- <property_2>: only_during_admission
- <property_2_value>: NULL

Replacing the variables in the template with real values results in the following MySQL statement:

\[
\text{INSERT INTO seizure_type_semiologies(} \text{id, seizure_type_semiologies, modifier, only_during_admission)} \\
\text{VALUES (‘TSXP606170783305’, 40, ‘Left Arm’, NULL)};
\]
3.2.6 Data Auditing Measures

Since CSR prospectively collects patient data from multiple sites, it is important to have data auditing measures to audit the quality of data being integrated. We adapt two commonly used data quality measures, completeness and consistency, to facilitate the CSR data integration and curation process (refer to [20] for a review of data quality measures). We audit the data completeness according to the types of data (phenotype and EEG signal) and common data elements, respectively. To perform auditing for types of data, the completeness is determined by whether both phenotypical and EEG signal data for a patient have been uploaded. For this, we develop an interactive reporting interface for CSR data manager to audit the uploading status of phenotypical and EEG signal data from individual clinical sites to the CSR central repository. To perform auditing for each common data element, the completeness is calculated or measured by the number of patient visits with actual data for the common data element divided by the total number of patient visits. Note that one patient may have multiple patient visits.

Moreover, we monitor coding consistencies for categorical and ontological common data elements, which is calculated by the number of valid values for the common data element divided by the total number of its actual values (i.e., the percentage of the valid coding values among all the actual values). For numeric common data elements, we use representation consistency, which is obtained by the number of numeric values divided by the total number of its actual values (i.e., the percentage of the valid numeric values among all the actual values).

4 Results

The CSR data integration pipeline and web-based data curation system have been implemented and deployed at https://medcis.case.edu/medcis. A collection of 95 common data elements were identified by CSR domain experts. A total of 629 patients have been recruited from seven clinical sites from October 2014 to February 2016. Among them, 393 are from UH, 65 from NYU, 11 from UCLA, 54 from NW, 44 from TJU, 56 from UCL, and 6 from UIowa. These numbers were generated in the reporting interface shown in Figure 5 (the numbers in the parentheses after the site names). The EEG signal data associated with these recruited patients exceeded 7TB. Since CSR is an ongoing effort, more patient data will be integrated and curated.

4.1 Web-based Data Curation Interface

Figure 4 presents a screenshot of the web-based curation interface for the section CLASSIFICATION OF PAROXYSMAL EPISODES. Clicking the “Edit” button for a common data element such as Semiology will trigger an interactive...
dialogue for a curation widget (see the red arrow in Figure 4). Clicking the “Update” button when the curation is done will save the curation result to the backend database and reflect the updated result in the web interface. Such web-based data curation widgets not only provide domain experts or data curators an intuitive interface to review and correct data, but also ensure the quality of the corrected data by strictly conforming to the standardized codings used for data capturing in OPIC. The CSR domain experts have been actively using the data curation interface to correct unintended errors and inconsistencies found between patient phenotypic information and EEG signal data.

4.2 Data Auditing Measures

Figure 5 shows the interactive reporting interface for auditing the completeness of the types of data uploaded from multiple sites. Domain experts can configure the report results by choosing sites of interest and the upload status of different types of data for each patient (EEG denotes the EEG signal data, PDF denotes the phenotypic data in PDF report). The returned results also include the number of visits for each patient (or Study ID). The customized reports can be exported to CSV files for the data manager in the CSR central repository to track down the types of missing data for each site and each patient.

<table>
<thead>
<tr>
<th>Select Sites</th>
<th></th>
<th>Upload Status</th>
<th>No of Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>UH (305)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYU (59)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCLA (11)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NW (54)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TJU (44)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCL (62)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIowa (6)</td>
<td>EEG,PDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Select Upload Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EEG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the data completeness of 10 common data elements by different sites and in total dated to February 2016. Age, Gender, and Drug achieved over 90% completeness. For Age, a total of 97.76% data completeness was obtained considering all sites. Individually, UH achieved 97.6% (487/499), NYU and UCLA got 100%, NW got 96.3%; but TJU, UCL, and UIowa only had 0% completeness for Age. The reason of 0% completeness for these three sites is that the OPIC instances deployed in these sites are not up-to-date, which is a common issue for multisite data integration or federated query (referred as data release cycle synchronicity in [19]). The most up-to-date percentages will be immediately available after the OPIC instances are updated.

For data consistency, all categorial and ontological data elements achieved 100% coding consistency, which demonstrates the effectiveness of using standardized codings for data capturing and integration. In addition, all numeric data elements achieved 100% representation consistency. Such data quality measures have been incorporated into the curation system in real time to facilitate data integration and curation.

5 Discussion

This paper presented ODaCCCI, an ontology-guided data curation system for supporting multisite data integration in CSR. It is adaptable to other multisite clinical data integration and curation workflows because of its general design
Table 2: Data completeness in terms of 10 common data elements for multiple sites up to February 2016.

<table>
<thead>
<tr>
<th>CDE</th>
<th>UH</th>
<th>NYU</th>
<th>UCLA</th>
<th>NW</th>
<th>TJU</th>
<th>UCL</th>
<th>Iowa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>97.6%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>96.3%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>97.76%</td>
</tr>
<tr>
<td></td>
<td>(487/499)</td>
<td>(67/67)</td>
<td>(6/6)</td>
<td>(52/54)</td>
<td>(0/0)</td>
<td>(0/0)</td>
<td>(0/0)</td>
<td>(612/626)</td>
</tr>
<tr>
<td>Gender</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>3.7%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>92.52%</td>
</tr>
<tr>
<td>Drug</td>
<td>93.39%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>81.48%</td>
<td>100.0%</td>
<td>50.0%</td>
<td>100.0%</td>
<td>92.09%</td>
</tr>
<tr>
<td></td>
<td>(466/499)</td>
<td>(67/67)</td>
<td>(6/6)</td>
<td>(44/54)</td>
<td>(40/40)</td>
<td>(12/24)</td>
<td>(5/5)</td>
<td>(640/695)</td>
</tr>
<tr>
<td>Semiology</td>
<td>78.76%</td>
<td>100.0%</td>
<td>16.67%</td>
<td>74.07%</td>
<td>92.5%</td>
<td>54.17%</td>
<td>80.0%</td>
<td>79.86%</td>
</tr>
<tr>
<td></td>
<td>(393/499)</td>
<td>(67/67)</td>
<td>(1/6)</td>
<td>(40/54)</td>
<td>(37/40)</td>
<td>(13/24)</td>
<td>(4/5)</td>
<td>(553/695)</td>
</tr>
<tr>
<td>Etiology</td>
<td>90.58%</td>
<td>73.13%</td>
<td>16.67%</td>
<td>37.04%</td>
<td>40.0%</td>
<td>100.0%</td>
<td>75.97%</td>
<td>79.71%</td>
</tr>
<tr>
<td></td>
<td>(452/499)</td>
<td>(67/67)</td>
<td>(1/6)</td>
<td>(40/54)</td>
<td>(40/40)</td>
<td>(5/5)</td>
<td>(5/5)</td>
<td>(554/695)</td>
</tr>
<tr>
<td>EEG Type</td>
<td>90.38%</td>
<td>16.42%</td>
<td>100.0%</td>
<td>88.89%</td>
<td>12.5%</td>
<td>12.5%</td>
<td>80.0%</td>
<td>75.97%</td>
</tr>
<tr>
<td>Epileptogenic Zone</td>
<td>73.95%</td>
<td>88.06%</td>
<td>16.67%</td>
<td>51.85%</td>
<td>47.5%</td>
<td>50.0%</td>
<td>60.0%</td>
<td>70.65%</td>
</tr>
<tr>
<td></td>
<td>(369/499)</td>
<td>(59/67)</td>
<td>(1/6)</td>
<td>(28/54)</td>
<td>(19/40)</td>
<td>(12/24)</td>
<td>(3/5)</td>
<td>(491/695)</td>
</tr>
<tr>
<td>MRI/CT status</td>
<td>65.73%</td>
<td>79.1%</td>
<td>66.67%</td>
<td>85.19%</td>
<td>92.5%</td>
<td>37.5%</td>
<td>100.0%</td>
<td>69.35%</td>
</tr>
<tr>
<td></td>
<td>(328/499)</td>
<td>(53/67)</td>
<td>(1/6)</td>
<td>(46/54)</td>
<td>(37/40)</td>
<td>(9/24)</td>
<td>(5/5)</td>
<td>(482/695)</td>
</tr>
<tr>
<td>Ictal Seizure Type EEG</td>
<td>62.73%</td>
<td>85.07%</td>
<td>100.0%</td>
<td>64.81%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>65.65%</td>
</tr>
<tr>
<td></td>
<td>(313/499)</td>
<td>(57/67)</td>
<td>(1/6)</td>
<td>(35/54)</td>
<td>(0/0)</td>
<td>(0/0)</td>
<td>(0/0)</td>
<td>(411/626)</td>
</tr>
<tr>
<td>Epileptiform Discharge</td>
<td>57.31%</td>
<td>77.61%</td>
<td>16.67%</td>
<td>57.41%</td>
<td>8.33%</td>
<td>80.0%</td>
<td>57.99%</td>
<td>57.99%</td>
</tr>
<tr>
<td></td>
<td>(286/499)</td>
<td>(52/67)</td>
<td>(1/6)</td>
<td>(31/54)</td>
<td>(27/40)</td>
<td>(2/24)</td>
<td>(4/5)</td>
<td>(403/695)</td>
</tr>
</tbody>
</table>

of defining common data elements and ontology-based vocabularies, and dynamic generation of web-based curation widget and dynamic interaction with backend databases.

For CSR, common data elements show different data completeness for disparate sites (Table 2). Since it is at the early stage of patient recruitment and data collection, values for certain data elements may still be missing from the source sites. Realtime data auditing from the CSR central curation system enables the center data curators to monitor the data entry process for each site, so that they can make data entry suggestions to each site for data elements with low completeness.

A limitation of ODaCCI is that it currently does not automatically handle the case that different sites may have data about the same patient. This is because OPIC instances are distributively deployed in multiple sites and cannot check potential duplicated patient information across different sites. Manual effort is needed to identify and resolve duplicated cases. Another limitation is that the web-based user interface has not been formally evaluated by end users. We plan to perform a usability evaluation of the web-based data curation interface and data auditing interface.

**Related Work.** Several related research efforts have been focused on multisite clinical research data integration and sharing. For instance, the Shared Health Research Information Network (SHRINE [21, 22]) was developed as a general clinical data integration system with federated query that could aggregate patient observations from multiple hospitals. It has been implemented for multisite studies of autism co-morbidity, colorectal cancer, diabetes, and others. SHRINE’s core ontologies include ICD-9-CM for diagnoses, RxNorm for medications, and LOINC for lab tests. The Human Studies Database Project (HSDB [23, 24]) has developed an informatics infrastructure for federated access and query of human studies databases. HSDB also developed the Ontology of Clinical Research (OCRe) to facilitate the data sharing. Similar to SHRINE and HSDB, ODaCCI uses a domain ontology to facilitate data integration.

There are also work on data quality assessment (DQA) of multisite clinical data. In [3], a “fit-for-use” conceptual model was proposed for DQA with a process model for conducting multisite DQA based on EHR data. In [25], DQA was performed based on EHR data from multiple student health centers in the College Health Surveillance Network. In [26], DQA was focused on Haiti’s national electronic medical record (EMR) systems, with an interactive “DQ dashboard” developed for quality improvement. Distinct from these studies [3, 21, 22, 23, 24, 25, 26], which focused on either multisite data integration or DQA, ODaCCI handles both multisite data integration and DQA. In addition, our web-based data curation system enables correction of data errors after data has been integrated concurrently.

**6 Conclusion**

This paper presented ODaCCI, an ontology-guided data curation system to address the data quality assurance needs for ongoing multisite clinical research data integration in the Center for SUDEP research. Phenotypic data and EEG signal data of 629 patients have been acquired and integrated from seven clinical sites between October 2014 and February 2016. Different types of data auditing measures have been incorporated in real time to help the CSR domain
experts with their curation work. ODaCCI supports continuous and asynchronous data quality improvements while data has been ingested and integrated from multiple clinical sites.

Acknowledgement

This research was supported by the NINDS-funded Center Without Walls for Collaborative Research in the Epilepsies U01 awards (U01NS090408 and U01NS090405), as well as the University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR000117).

References

Nine Principles of Semantic Harmonization

James A. Cunningham¹, Ph.D., Michel Van Speybroeck, M.Sc.², Dipak Kalra³, Ph.D., Rudi Verbeeck, Ph.D.²

¹Health e-Research Centre, The University of Manchester, Manchester, UK; ²Janssen Pharmaceutica, Beerse, Belgium; ³The European Institute for Health Records (EuroRec), Sint-Martens-Latem, Belgium

Abstract

Medical data is routinely collected, stored and recorded across different institutions and in a range of different formats. Semantic harmonization is the process of collating this data into a singular consistent logical view, with many approaches to harmonizing both possible and valid. The broad scope of possibilities for undertaking semantic harmonization do lead however to the development of bespoke and ad-hoc systems; this is particularly the case when it comes to cohort data, the format of which is often specific to a cohort’s area of focus. Guided by work we have undertaken in developing the ‘EMIF Knowledge Object Library’, a semantic harmonization framework underpinning the collation of pan-European Alzheimer’s cohort data, we have developed a set of nine generic guiding principles for developing semantic harmonization frameworks, the application of which will establish a solid base for constructing similar frameworks.

Introduction

The quantity, quality and prevalence of digitally represented medical data available for research world-wide is now such that a tipping point has been reached in terms of the ability of that data to inform novel and innovative methods of research¹. This so called ‘Big Data’ revolution, where the volume of data available is such that previously unobtainable research results are obtainable will inevitably inform medical research practice moving forwards². That data is universally available in digital form does not however mean that it is available and represented in the same form – data sets large enough to catalyze novel research will inevitably come from a myriad of sources³. These sources, at least in the present, can and will draw from a large number of potential methods of representing data. As such, drawing data from multiple sources and combining it into a form upon which research can be conducted will require a process of either combining data into a single representation or explicitly mapping and translating between items of knowledge from the various sources⁴.

Semantic harmonization then is the process of combining multiple sources and representations of data into a form where items of data share meaning⁵. Harmonized data imparts the ability of allowing single given questions to be asked and answered across the data as a whole, without need to modify or adapt queries for a given data source, invaluable as a tool for researchers. The process of harmonizing data is far from trivial though. The technical undertaking of specifying a representation for capturing harmonized knowledge, the person-effort of specifying domain specific mappings between instances of data and the fact that for any given items of knowledge from within the same domain there may not be a universally agreed translation between them combine to make the task of harmonizing data, even for relatively specialized and small domains of knowledge, onerous if not impossible. Regardless of the feasibility of successfully harmonization any given data sets, undertaking the process and building a framework for semantic harmonization is a common and perhaps inevitable part of projects leveraging ‘Big’ medical data⁶.

Given that the implementation of some form semantic harmonization framework will necessarily form part of a project that utilizes medical data from across multiple sources, it seems natural to ask the question as to what common frameworks are available for enabling these harmonization processes? In answering this question we will draw a distinction between a method of harmonization and a framework for harmonizing data. A harmonization method is the underlying technical or logical means of specifying knowledge or mapping between items of knowledge, Formalisms such as ontological representations and their related reasoning mechanisms⁷, or openEHR archetypes⁸ would qualify as harmonization methods. It is uncommon and perhaps unwise for projects dealing with semantic harmonization to go down the path of specifying novel underlying representation formats. A harmonization framework on the other hand would be the overarching infrastructure for enabling harmonization and utilizing its
results as well as the specific choice of how to use a chosen harmonization method. While systems such as I2B2 would fall into the category of harmonization frameworks it is still often necessary to develop novel frameworks for capturing data, particularly where the domain the data is being drawn from is focused or specialized. One such area would be that of cohort data. The collection of cohort data in a particular disease area will often focus on items or forms of knowledge specific to that area and the representation and capturing of that knowledge may require methods and formalisms unique to the domain in question.

Regardless, the development of frameworks for enabling semantic harmonization, particularly given the ongoing development of the field of ‘Big Data’ research in the medical informatics space, will continue. Each such system will be different; they will take varying approaches to how to harmonize data and how to enable the use of that data. We believe however that there should be a set of core principles that underlie the development of any such system. Key to these principles are features that inform the separation of concerns in the representation of knowledge and approaches to the stratification of knowledge. These are outlined following a description of the project that informed the development of these principles.

The EMIF Knowledge Object Library

The European Medical Informatics Framework (EMIF) is a large-scale collaboration between academic research institutions, small and medium sized companies and members of the European Federation of Pharmaceutical Industries and Associations (EFPIA), part of the Innovative Medicines Initiative (IMI). It aims to develop a suite of tools for enabling the reuse and linkage of healthcare data from across Europe. Specific use cases within EMIF are drawn from the participation of data cohorts from the Alzheimer’s and Metabolics spaces. It is with a focus on the Alzheimer’s space that we have produced the EMIF Knowledge Object Library (EKOL). This is a system for capturing both specific local knowledge from Alzheimer’s disease patient cohorts and mapping this knowledge into a global representation that allows the combination of data from multiple cohorts into a single representation, giving researchers the ability to launch research queries against this unified view of the data. Data recorded by the participating cohorts is universally represented in individualized data schemas and the majority of the data is coded with reference to local vocabularies, when formally coded at all. This led to the need for a system that dealt very explicitly with local data in a form unique to its source, but local data that in turn generally referred to the same set of global concepts.

One practical outcome of this work was the application of the Knowledge Object Library as the underlying mechanism used to capture, process and provide data for the EMIF ‘Participant Selection Tool’ (PST). The PST is a web based tool that presents researchers with a list of approximately 20 core criteria present in the majority of EMIF AD cohorts (covering items such as participant age, diagnosis and mental state examination scores), showing counts of patients per cohort that have records covering those items of data. It then allows researchers to select logical combinations of these criteria and their values and dynamically adjusting the count of patients that match the selected criteria. So, for example, we could select ‘Patients over 50 with a diagnosis of Mild Cognitive Impairment OR Patients over 60 with a diagnosis of Alzheimer’s Disease’ and be presented with an updated count of the number of patients across the EMIF cohorts that match this selection. An illustrative screenshot of the PST is shown in figure 1. The PST is part of a wider tool-chain that ultimately allows for the specification of research studies and the export of associated harmonized data for use by researchers. The information presented by the PST comes from data harmonized using the EKOL harmonization framework.

As part of the process of specifying, designing and developing the EKOL system we explicitly drew out and developed a set of principles for designing semantic harmonization frameworks. These were developed through a combination of analysis of user requirements, an iterated agile approach to the development of the associated software which allowed for philosophical design decisions to be rapidly tested and potentially rejected in real-world settings and analysis of existing systems. These principles, outlined in the next section, both informed the development of the EKOL framework and directed the specification and design and application of the Participant Selection Tool.
The successful harmonization of data for use within the PST and its deployment as a tool for research offer an empirical justification that the principles we explicated during the development of the EKOL framework at least offer the beginnings of a generalizable approach to the design and development of similar systems. These principles are outlined below.

**Principles of Semantic Harmonization**

Our work in the development of EKOL has lead to the formulation of the following set of generic guiding principles for developing frameworks for semantic harmonization. Given the range of ways in which knowledge can be defined, represented and captured, we use the generic term *knowledge object* to describe a singular representation of an item of knowledge, independent of the underlying formalism for capturing that item of knowledge. In this generic sense it is the representation of knowledge objects combined with the ability to the system to capture and specify them that underlies the functionality of any semantic harmonization system.

The nine points, presented below, when taken together are designed to guide the development of frameworks for supporting and enabling the semantic harmonization of medical data from multiple sources.

1. **Separate technical from semantic harmonization.**

Data can be stored in many different file formats or databases. The ability to access and process available data is clearly an essential prerequisite to being able to semantically harmonize that data. We refer to the process of transforming data into the same technical storage implementation or transferring it to a system that can be queried by the same standard query engine as *technical harmonization*. The process of making all the data available on compatible platforms is a technical problem and should be separated from the semantic harmonization of the data, which is a data content problem - *technical harmonization is not semantic harmonization*

As such:
- A technical connector to the local data source should be developed only once.
- Technical upload scripts or tools should not be impacted by a change in semantics, such as a change in a vocabulary.
- Conversely, changes or additions to technical harmonization infrastructure or process should not impact semantic modelling or representation.

2. **Distribute ownership of local and global knowledge objects.**

In general, it is the role and responsibility of a subject matter expert to specify how variables need to be harmonized semantically. Semantics encapsulate meaning and ultimately it is only a subject matter expert, particularly in the medical domain, who can specify such meaning. However, knowledge of how variables and data can be utilized, particularly for research, may fall within another domain of knowledge, such as that of the researcher, whose knowledge and expertise, whilst complimentary to the domain expert, is not necessarily the same. They should be able to describe the variables they require and, if applicable, how normalization or harmonization should be performed. Variables that are used for data analysis are the global representations of knowledge. They are defined independently of any data source. Knowledge of measurement protocols and local variables – how variables are measured and what they really represent – are, on the other hand, available only at the data source. It is therefore the responsibility of the data source custodian to describe these variables semantically and create a local representation of knowledge for each variable they want to make available.

Bridging the gap between local knowledge, as described by local experts, and global knowledge, as requested by researchers, is a joint responsibility and effort and is part of the workflow of a semantic harmonisation process. As such:

- Ownership of local and global representations of knowledge objects should be distributed according to the location of expertise.
- Explicit distinction should be made between what is local knowledge and what is global.
- It should be possible to build up and maintain a library of global knowledge objects over time – both global and local concepts should be extensible.
- Global knowledge should make only generic reference to local knowledge, but local knowledge should be grounded in terms of global concepts.

3. **Separate vocabulary from structure.**

Data integration in the medical field is often achieved by the use of common vocabularies or taxonomies. Examples of well-known taxonomies are SNOMED-CT\(^\text{[12]}\) or the NCI thesaurus\(^\text{[13]}\). CDISC\(^\text{[14]}\) also specifies vocabularies that need to be used for naming variables or for standardizing values for those variables used commonly in clinical research.

Similar to the distinction we have drawn between technical and semantic harmonization, we will also highlight a distinction between the act of specifying a vocabulary for grounding the meaning of items of knowledge via explicit reference to vocabularies and the act of structuring items of knowledge such that their internal relationships carry meaning. Whilst structure is often present to varying degrees within formal medical vocabularies and this structure can be leveraged in the formalization of semantic harmonization frameworks, drawing a conceptual distinction between vocabulary and structure can avoid a system relying on the tacit knowledge carried in language to derive its full utility. When the meaning of knowledge is externalized a system can no longer manipulate that knowledge and utility is lost.

So, the concept of vocabularies should be incorporated into the underlying representation of knowledge, since without any reference back to the real world the ultimate meaning of the semantics of knowledge is lost. However, a knowledge object should be a general construct that gets its definition from a link to a vocabulary term, rather than begin inferred from its structure.
As such:

- Knowledge representation structures should be generic, with terminology details and dependencies separated by reference to vocabularies
- Knowledge representations should retain meaning across vocabularies, and maintain meaning and structure in the absence of a vocabulary
- The addition of new vocabularies should be possible with zero, or minimal, change to the structure of global knowledge representation

4. **Re-use standard vocabularies where possible.**

Following on from point 3, we note that many medical vocabularies exist today and a semantic harmonization framework should re-use what is available. This both removes the need to maintain vocabularies and, given the ability to take on board new vocabularies or extend existing ones. The design of a novel semantic harmonization framework does not necessitate the design of a novel vocabulary, these are separate (though clearly related) areas of research, and should not be conflated.

As such:

- Existing vocabularies should always be used in the first instance
- Where there is a need to define new terms not available in a public vocabulary there should be mechanisms in place to accommodate this.
- Such mechanisms should be used sparingly

5. **Use declarative mappings.**

Domain experts are generally better at describing the logic of a mapping between knowledge objects, rather than describing the control flow of the computation of that mapping. Given the need to capture knowledge from domain experts, rule based rather than computation-based systems are better. A domain expert, assuming the role knowledge author, needs to be presented with a mechanism for capturing rich and complex domain knowledge. Declarative rule based systems are best suited for this purpose.

As such:

- Mappings between knowledge objects should be described using rules.
- Mapping rules need to be unambiguous and executable and as such expressed in a formal language
- Such rules should still be simple enough to explain in plain language to other domain experts for verification.

6. **Code isolation.**

The use of rules to specify mappings has the added benefit that a rule can be specified as a stand-alone object, independent of other mapping rules. This contrasts to scripts or other representations of computational flow, where changes to the code can have an impact on code statements further down the control flow. A rule for specifying knowledge should encapsulate a single concept and should be as limited as possible in scope. With this approach rules can be replaced or amended without affecting other rules within the system and the effort of authoring and maintaining rules can be distributed and scaled.

As such:

- A given mapping rule should be assigned to or linked with a single knowledge object
- A rule should contain the instructions on how to calculate a given knowledge object only from upstream knowledge objects
- The scope of rule should be limited solely to the knowledge object it describes and the upstream knowledge objects it is derived from – it should be unaware of any other variables of knowledge objects.
- The impact of changing any given rule on the global structure of knowledge should be minimised.

7. **Enable integrated security and provenance.**

Semantic harmonisation more often than not will need to occur across data sources where the ownership of data (in terms of access control) differs between sources. It is essential that the data custodian for a given data source can control access to the data source variables or local knowledge objects, even when data is ultimately queried at the level of the global knowledge objects. Access to a local knowledge object should be propagated through the dependency graph to global knowledge objects. A knowledge representation system should have constructs in place that can handle this propagation.

Access to data can be restricted based on a number of dimensions:

- **Users**: access is restricted to certain users or research teams, based on the description of their scientific use case.
- **Variables**: the researchers should only have access to the data they need to answer the research question. To avoid having to set permissions on each variable, data source custodians can group variables (as is done for example in CDISC domains) or provide access to the complete data source.
- **Patients**: data source custodians could allow access to data of a subset of patients only, based on inclusion or exclusion criteria or on legal restrictions.
- **Time**: access to source data can be limited in time.
- **Permission level**: access could be granted to the patient-level data or to aggregated knowledge objects only. At the level of the outflow tools it could be restricted to pre-defined reports or could allow custom queries on the data.

As such:

- Items should be grouped as much as possible in order to minimize administrative complexity (for instance users can be grouped in teams, institutions, capability groups etc.)
- In the context of describing security variables should be grouped in domains or by data type (e.g. clinical data vs. high dimensional data).
- Patients should be grouped and categorized by originating data source
- Derived instance values corresponding to knowledge objects should always reference source data instances from which they have been derived, allowing full traceability of every derived value (data provenance).
- The data source of a derived value should be inferable directly from the mapping rules.

8. **Separate WHAT is measured from HOW the measurement is done.**

When mapping data from different data sources we clearly need to make sure that variables that are mapped to a common global knowledge object are compatible, i.e. that they measure the same thing. Using terms of the Ontology for Biomedical Investigations (OBI)\(^\text{15}\), they need to share the same “quality” that “inheres in” instances of the same “material entity”. If there were no measurement errors (instruments have infinite accuracy), variables that measure the same quality in the same material entity could directly be pooled.

However, the actual measured value is also determined by the measurement protocol and the accuracy of the instruments used. Compatibility of protocols or normalization of values (to factor out the effect of the protocol) is use case dependent. The scientist analysing the data needs to decide if data measured using different protocols can be pooled. Whilst it is not necessarily the case the full details of a given measurement protocol should be modelled,
we recommend specifying the independent variables of a protocol using vocabularies so they become comparable between data sources. Figure 2 gives a graphical overview.

![Diagram]

**Figure 2.** Compatibility of measurement variables depends on what is being measured and on how the measurement was performed (the protocol).

The knowledge object representation should be able to capture the “quality” and “material entity” for the variable it models, but should also capture as many of the protocol’s independent variables as deemed useful. Independent variables can be used as filters at query time, or be retrieved as additional information to the actual measurement value. They can also be moved into the “quality” to make the knowledge object more specific.

A typical example would be a knowledge object for hippocampal volume. The “quality” in this example is “volume”. The “material entity” is the hippocampus (e.g. bilateral). But there are many independent variables in the measurement protocol, such as the type of scanner, manufacturer of the scanner, image acquisition sequence, field strength, segmentation algorithm, digital brain atlas used etc. We could also define a knowledge object for “hippocampal volume as measured on 1.5T MRI using Freesurfer cross-sectional”, which would fix some of the independent variables. Such a knowledge object would measurements of different data sources more comparable but decreases the number of data sources that can deliver the data represented by the knowledge object.

As such:

- A knowledge representation should always distinguish what is measured from how it is measured
- This principle should be applied to measured observations both for continuous and categorical variables, textual information and other coded observations.

9. **Balance generic versus specific descriptions.**

The range of possible variables that can be captured by knowledge objects is very large. For clinical data common use cases at least cover all the domains described in the CDISC SDTM standard. CDISC models each domain differently and specifies columns that are common to (most) domains, but also lists columns that are domain specific. Most CDISC domains specify one column per variable to record measurement values (pivoted data format). The demographics domain, for example, has a separate column for Age, Sex, Race etc. But some domains use one column to record the test that was performed and another for the measured value (unpivoted data format). The lab domain, for example, uses the column LBTEST for the laboratory test name and LBORRES for the (original) measured value. Other common data models such as OMOP take a more general approach and store all clinical observations in a single database table (Observation). Patient related information, however, is also stored in tables such as Person or Condition_occurrence.
Choosing specific data structures for different types of variables can customize the information that is recorded. A generic data structure may have difficulty recording domain specific details. However, storing information across different tables and columns can result in domain dependent models.

There is no clear guideline on what the optimum generality level of a data model should be. The structure of a knowledge object tries to be sufficiently general to be able to capture the anticipated types of data and its metadata, but over-generalization has the drawback of a possible loss of domain dependent detail.

This challenge is equally recognised in the development of semantic interoperability sources such as detailed clinical models for healthcare information exchange. Discussions about how to balance generic versus very specific clinical models are currently ongoing within the Clinical Information Modeling Initiative (CIMI). There is a complementary ongoing debate about the level of detail and completeness that models (and knowledge objects) should aim to specify: should one aim for large very comprehensive representations or smaller fragments that may be combined in different ways.

This final principal is in a sense necessarily vague: in developing a semantic harmonization framework, the representation of knowledge into which data is harmonized needs to be generic enough to harness the benefits of simplicity without losing the ability to capture domain specific structure that can come from a bespoke knowledge representation structure.

As such:

- The design of a knowledge representation scheme for semantic harmonization should be as generic as possible.
- Specificity of representation should only occur where the representation of that knowledge cannot be generalised, i.e. where there are features of the representation not shared by any other objects in the general knowledge representation scheme.

Discussion

Systems or frameworks for harmonizing knowledge from multiple heterogeneous data sources must, where knowledge is being explicitly represented, be designed around that representation of knowledge. In this sense it is crucial that a clear, principled approach to the design of those systems, particularly when it comes to the stratification, specification and codification of knowledge, is taken. In this paper we have outlined a set of guiding principles that we believe when followed will lead to the design of frameworks for semantic harmonization that are robust, future proof and able to capture the full richness of knowledge within a given domain.

The nine principles outlined above can be applied en masse or adopted piecemeal, and conflicting views or implementations are inevitable. We do believe however that to a large extent the nine principles express a clarification of common design philosophy rather than a dictation of choice of approach. The value in outlining the set of principles comes from their explication rather than from the authors having made binary choices on behalf of others.

The focus of the principles outlined here focus specifically on areas of semantic harmonization that address aspects of conceptual modeling as opposed to value modeling (in terms of the ISO 11179 MDR standard at the data element concept rather than data element value level). This focus stemmed primarily from the emphasis of this work being driven by the need to harmonize cohort data, where we found that the primary differences between data sources lay in the structuring rather than the value of measurements and results. Adaptation or revision of these principles to account more for value driven harmonization efforts could be an important extension of this work.

We have drawn a distinction in this work between both between methods and frameworks of harmonization and between technical and semantic harmonization and believe that the field as a whole would benefit from applying further similar analysis to the structure of the field itself and particularly application development within it. The field of software development has benefited immensely from a rich and varied analysis of the types of approaches that can be taken to software development itself, leading to myriad approaches to the ‘art’ of the field. We believe that the area of semantic harmonization would benefit similarly from a similar reflective approach being taken.
Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement n° 115372, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kind contribution.

References

A K-Reversible Approach to Model Clinical Trajectories

Filip J. Dabek, MSc¹, Jesus J. Caban, PhD¹
¹National Intrepid Center of Excellence, Walter Reed National Military Medical Center, Bethesda, MD

Abstract

A clinical trajectory can be defined as the path followed by patients between an initial health state \( s_i \) such as being healthy to another state \( s_j \) such as being diagnosed with a specific clinical condition. Being able to identify the common trajectories that a group of patients take can benefit clinicians at identifying the current state of patient and potentially provide early treatment to avoid going towards specific paths. In this paper we present our approach that enables a clinical dataset of patient encounters to be clustered into groups of similarity and run through our algorithm which produces an automaton displaying the most common trajectories taken by patients. Furthermore, we explore a dataset of patients that have experienced mild traumatic brain injuries (mTBI) to show that our approach is effective at clustering and identifying common trajectories for patients that develop headaches, sleep, and post traumatic stress disorder (PTSD) post concussion.

Introduction

A clinical trajectory can be defined as the path followed by patients between an initial health state \( s_i \) such as being healthy to another state \( s_j \) such as being diagnosed with a specific clinical condition. Commonly clinicians review a patient’s medical history to better put the medical findings within the context of the patient. While collecting and reviewing medical history is essential to providing personalized treatment, current clinical decision support systems are not effective at aggregating and understanding how a group of patients go from state \( s_i \) to another state \( s_j \). Therefore, being able to identify the common trajectories that a group of patients take can benefit clinicians at identifying the current state for a particular patient and potentially providing early treatment to avoid going towards some specific paths. Despite the significant amount of longitudinal information that Electronic Health Records (EHRs) include related to patients’ clinical encounters, determining the most frequent clinical trajectories followed by a given group of patients is still a challenging task. Often within a given cohort of patients, a group of patients take similar paths to a certain disease while others take a vastly different path. Due to the increased interest in understanding how a patient’s condition will develop over time, understanding these varying paths can benefit healthcare by providing physicians and patients crucial information in preparation for the future. Clinical trajectories possess the power of providing physicians and patients with a visual representation that can easily be understood and analyzed. They have the potential to uncover hidden information in data that could not be seen otherwise. With these goals in mind, in this paper we present a framework for creating meaningful clinical trajectories using longitudinal clinical data. First we describe the challenges faced by researchers in using clinical data, next we describe some of the previous work, then we present our approach to cluster patients and to build a condensed model that can be analyzed, followed by discussing how we’ve applied our approach to a large clinical dataset, and finally we conclude the paper and describe some of the future work.

Healthcare data has been predicted to exceed 25,000 petabytes in 2020 compared to 500 petabytes in 2012, an increase by a multiplier of 50¹. With this rise in the amount of available big data in healthcare, there exists an opportunity to apply machine learning techniques on a large scale to identify key information in patient trends. Understanding what causes patient deaths, what is effective at improving a patient’s life, and how a population of patients end up in a certain diagnosis are just a few examples of the insightful analytics that can be discovered in this data.

However, even though there are many points of information that are desired to be studied, the vast amount of data provided along with the many different types of patients poses extreme challenges to researchers. Not only are efficient algorithms and powerful systems required to extract and perform computations, but understanding the wide variety of patients is key in identifying breakthroughs in healthcare research to assist future patients. It may be easy for a physician to analyze one individual patient to understand their medical history, but at a large scale of data it is not feasible to analyze each patient individually in an attempt to get an overall view of the population. The amount of time it would take to analyze, with more data being added by the minute, would not provide tangible benefits for the short term. Therefore, methods that attempt to combine similar patients and produce a model of the patient population...
will prove to be effective at providing physicians with a condensed representation of the population in order to assist in preventative care and supporting future patients that will undergo similar conditions.

**Background**

Research in the area of clustering big data has proposed methods in which an ensemble model of statistics information and word sense information is used to cluster documents, a framework in which the popular k-means algorithm applied to a weighted linear co-association matrix to cluster biomedical data consisting of text and images, and applying data clustering algorithms to the problem of big data. All of these approaches have shown the benefit that clustering has: the power to be able to identify groupings in data that otherwise could not be found with alternate methods.

In the clinical realm, approaches to understanding the clinical trajectory of a patient, by means of clustering, have focused on a specific disease with the intention of being able to predict whether a new patient will develop a similar diagnosis or not. In the realm of PTSD patients, Bryant et al. studied the long-term trajectory of PTSD patients over 6 years and classified them in chronic, recovery, worsening/recovery, worsening, and resilient groupings. The study found that analyzing patients over time provides a more accurate means to identifying and predicting PTSD rather than relying on hospital admission analysis. A similar study analyzed patients 12 months after a burn and also found four different trajectories for patients with PTSD. In addition, it was found that the risk factors differed between trajectories indicating that each group possessed similar traits which could be used in clinical practice. Another study concentrated on children after an accidental, but traumatic, injury in which they developed Post Traumatic Stress Symptoms (PTSS). In this study the researchers utilized group-based trajectory modeling to identify patterns of PTSS and found three distinct trajectory groups where pre-injury risk factors were predictive of the corresponding group. With this information the researchers concluded that identification of distinct trajectory groups can help understand the course from traumatic injury to PTSS and the necessary treatment for a child. Gotz et al. utilized patient similarity metrics to visually analyze patient clusters. In addition, various Bayesian and mathematical models have been utilized to build disease models and cluster patients, but lack the ability to provide a simple visualization of disease trajectory. Furthermore, several Harvard researchers have shown their ability to classifying and predicting long-term medication adherence using group-based trajectory models. All of these studies alike indicate that patients’ trajectories can be grouped together to find a common path and give clinicians an insight into the expected diagnosis path and optimal treatment.

**Approach**

For our approach we attempted to classify the patients into their respective groups, based on the trajectory that they underwent, followed by representing each group’s trajectory with the most common paths highlighted in a visual manner using automata. Below we will outline the steps taken for each aspect of our approach.

**Clustering**

With the knowledge that not all patients are the same and that patients take varying, but similar paths, we used the popular clustering algorithm, k-means clustering, to identify the various groups. The k-means algorithm takes a set of feature vectors and a value of k which it then attempts to find k unique groups using a distance metric, in this case the Euclidean distance.

Our first attempt at clustering the patients was to associate each diagnosis with an integer and thus create a feature vector of the diagnoses in order. However, because patients have varying number of diagnoses and k-means requires uniform length, the algorithm ultimately clustered the patients based on the length of their original trajectory. With this we had to find an alternate method that did not discriminate between the patients based on their varying trajectory length. Using our previous work in which we predicted patient outcomes based on their medical history, from Electronic Health Records (EHR) data, we created a sparse matrix representation of each patient which we will briefly describe next.

To create meaningful input for the k-means algorithm, first the encounters / diagnosis tuples were transformed into a sparse matrix where each row was a diagnosis and each column corresponded to a different time point.

With the sparse matrix defined, we used the Bayesian Information Criterion (BIC) and the elbow criterion method.
to identify the optimal number of $k$ as this method identifies the point at which additional clusters would result in overfitting.

**Model Automata**

In an attempt to design a model to estimate the common trajectory followed by most patients from their initial injury to the first diagnosis of PTSD, we model the longitudinal clinical encounters as an automata and employ grammar induction algorithms to minimize the complexity of large automata. This approach allows us to simultaneously consider the path of N patients and minimize the graph into a single automaton that represents the common path. In order to understand our approach, we will first give a brief overview of automata.

Automata are self-operating machines that consist of states and transitions, where the input is compared against the transitions in order to move between states in the machine. Two forms of machines exist in finite automata: deterministic (DFA) and non-deterministic (NFA). For a given state and input symbol, deterministic machines only have one possible transition, whereas non-deterministic machines can have multiple transitions.

Using the two forms of automata, nondeterministic (NFA) and deterministic (DFA), will allow us to visualize the trajectory of patients over time. Both of these approaches represent the data in a unique way allowing for us to evaluate different aspects of the trajectories. It should be noted that both the NFA's and DFA's were run through a simple minimization algorithm that reduced the number of nodes for each individual patient trajectory. For example, a patient that followed the path of: \texttt{A -E1 -E2 -E3 -B -E4 -C -E5 -C -A} would be reduced to: \texttt{A -E1 -E2 -E3 -E4 -E6 -E5 -A}, as can be seen in Figure 1.

The first approach, using an NFA, treats each diagnosis as an individual node and stacks into a column representing an encounter. An example of a single patient’s NFA can be seen in Figure 2 where “f” corresponds to a Concussion, “P” corresponds to PTSD, “D” corresponds to Depression, and “E” corresponds to a diagnosis related to Speech. In this example we can see that the patient was diagnosed with P, D, and f in their first encounter, followed by P and D in their second encounter, etc. Representing a patient as an NFA allows us to understand the change in the number of diagnoses over time and represent the concept that a patient can take the path from any disease in a single encounter to any disease in the next.

The second approach, using a DFA, represents each encounter as a node with the diagnoses acting as paths leading into the node. This representation is shown for the previously presented patient in Figure 3, where the automaton looks to be smaller due to the lesser amount of nodes. This DFA approach will allow us to utilize a grammar induction algorithm for identifying a common trajectory amongst a group of patients.

Furthermore, both of these approaches can be augmented by placing a scale for the amount of days between encounters in order to understand the timeframe of the patient’s encounters better. The previous patient’s NFA and DFA were modified to display the amount of days between encounters in Figure 4.
Figure 3: A DFA representing six encounters corresponding to the path followed by a particular patient between his/her first concussion to PTSD.

Figure 4: Patient’s DFA (Top) and NFA (Bottom) with a timeline of days between encounters in the first row of each automaton. The timeline represents the number of days between mTBI and PTSD, which is 24 days in this patient’s case.

Using the previously demonstrated NFA and DFA’s of a patient’s trajectory, we applied a grammar induction algorithm to the automata with the goal of identifying the most common paths that a group of patients take within their trajectories. The specific algorithm that we chose was the K-Reversible algorithm as it treats all input as being equal, compared to other algorithms such as Gold’s Algorithm that requires positive and negative data.

Merging

A prefix tree acceptor (PTA) is a tree-like DFA built from a learning sample of strings $P = p_1, p_2, ..., p_n$ by converting all of the prefixes $p_i$ in the sample $P$ into states $Q = q_1, q_2, ..., q_n$, and constructing the smallest DFA that is a tree and consistent with the learning sample\textsuperscript{10}. For example, given a sample string set $P = aa, aba, bba$ it can be converted into a set of states $Q = q_1, q_2, ..., q_7$ as illustrated in Figure 5a.

One of the basic operations that can be performed on a PTA is a merging operation, which takes two states $(q_i, q_j)$ from an automaton and merges them into a single state\textsuperscript{16}. An example of the merging algorithm is shown in Figure 5b. The algorithm takes in two states, $q_i$ and $q_j$, that are to be merged together and then takes everything that points into $q_j$ and makes it point into $q_i$. In addition, everything that $q_j$ points to is now made to originate from $q_i$. This removes all of the transitions into and out of $q_j$ and transfers them to $q_i$, allowing the algorithm to now remove $q_j$ from the finite state machine.

Algorithms such as Gold’s Algorithm\textsuperscript{17}, RPNI\textsuperscript{18}, and K-Reversible\textsuperscript{16} start from a PTA and perform operations on it to try and create a DFA that recognizes the target language. By using these algorithms, we can input an automaton of clinical trajectories that will be merged leaving behind the most common paths taken by patients.

Minimizing the Automata

The specific algorithm that we chose was the K-Reversible Grammars Algorithm\textsuperscript{16,19} as it treats all input as being equal, compared to other algorithms such as Gold’s Algorithm that requires positive and negative data.

Figure 5: (a) Example of prefix tree acceptor built from strings $P = aa, aba, bba$. (b) Example of merging states $q_1$ and $q_2$. 

463
The K-Reversible Algorithm is classified as a look-ahead language that takes into account a length \( k \) sub-sequence of diagnoses at a time. This means that the algorithm starts at a state and looks backwards for up to \( k \) states, combining the diagnoses on the backwards path into a sequence. Subsequently, two states can be compared by contrasting their sequences of \( k \) length diagnoses. An example of this is shown in Figure 6, where states \( p \) and \( q \) are similar for \( k = 0 \) as “a” equals “a”, but for \( k = 1 \) or \( k = 2 \) they are not similar as “ad” does not equal “ab” and “adc” does not equal “abc”.

This K-Reversible algorithm was ultimately chosen for its ability to merge similar states based on a length \( k \) sequence of diagnoses as similar sequences of diagnoses can potentially lead to the same condition in a patient.

Breaking down the K-Reversible Algorithm, there are three cases that need to be considered. The first case looks for a state, \( p \), that has two identical transitions to two different states, \( q \) and \( q' \). Once this case has been matched then the states \( q \) and \( q' \) are merged together. This is shown in Figure 7a. The second case of the algorithm looks for two final states, \( q \) and \( q' \), that have identical paths leading into them of length \( k \), and merges them together. This is shown in Figure 7b. Continuing on with the algorithm, the subsequent states, \( p \) and \( p' \), are also merged. The third case of the algorithm looks for a state \( p \) that has two identical transitions leading into it from two different states, \( q \) and \( q' \), that also have an identical path leading into them of length \( k \). Once again, \( q \) and \( q' \) are merged together by this case. This is shown in Figure 7c.

Algorithm Modifications
We found that the three cases of the K-Reversible algorithm did not achieve an optimal automaton, through visual inspection, such that:

1. similar, but not identical, paths were not merged
2. duplicate paths were not removed
3. a patient’s path was merged with itself

With these flaws present, we modified the algorithm by adding three cases in addition to the original three cases.

First Modification The K-Reversible algorithm strictly considers paths without taking into account that two distinct paths may in fact be the same despite a difference in order of diagnoses. For example, there could be a state \( q \) that has paths of \([b, a]\) and \([a, b]\) leading into it, as shown in Figure 8a. In these types of occurrences, we assumed that the ordering of the diagnoses does not matter as both paths lead to the same state of the patient. This property of
the algorithm required a modification, accomplished by reordering the less probable path to be equivalent to the more probable path. Upon reordering and running the algorithm further, it will eventually merge these two paths into one.

**Second Modification** The second modification to the K-Reversible Algorithm was to remove identical paths to the same state, that was generally caused by the first modification. During the minimization of the grammar, the algorithm would cause there to be two states, p and q, where more than one path went from p to q with the same value as the transition. Therefore, by merging these paths, the grammar would be minimized to the furthest extent. An example of this modification is provided in Figure 8b, where the duplicate path of ‘a’ is merged into one single path.

**Third Modification** The third modification to the algorithm was with the goal of not allowing the algorithm to merge a patient’s trajectory into itself. What this means is that instead of treating a patient’s trajectory as one entire path, the algorithm would attempt to alter the trajectory to be minimal, resulting in only one node and the patient not being represented accurately. We therefore modified the algorithm to track the patient ID’s at each node and did not allow two nodes to be merged if the intersection of the ID’s contained a patient. This ultimately forced the algorithm to ensure that the diagnosis trajectory of each patient was treated as an individual sequence.

**Fourth Modification** Due to the large size of automata that can result from running the algorithm on an extensive dataset and due to the nature of clinical encounters: in that patients can take many varying, unique paths which cannot be merged together, it can become difficult to understand the common trajectories on a large automaton. This limitation in understanding resulted in us making one final modification which involved pruning the resulting automata to only consider the most probable/most taken paths. During the execution of the K-Reversible algorithm, we kept a count of the number of patients that took each path. Then, once the algorithm finished running we used the patient counts to compute the probability of each path being taken from a node, thus leaving us with a patient count and probability at each path. Using these two metrics, we were able to set varying probability and count cut-offs such that the resulting automaton would only include paths that were above the cut-off. This allowed for the size of the automaton to be reduced as can be seen in the next section.

With these modifications made to the algorithm we will explore, in the next section, the automata that result from running our modified algorithm on longitudinal clinical data.

**Application Domain**

For our results we will explore a dataset available to us centered around the application domain of patients that experience a concussion or a mild traumatic brain injury (mTBI) and the long term affects associated with this injury. Specifically, we will be looking at patients that develop diagnoses related to either headaches, sleep, and/or PTSD post concussion.

**TBI Data**

A concussion is a poorly understood mild traumatic brain injury (mTBI) that can alter the way the brain functions. During the last decade a significant amount of attention has been given to the acquisition of clinical data from patients suffering mTBI and psychological health (PH) problems after a concussion. The increased awareness has been in part driven by the Department of Defense (DoD), the National Football League (NFL), and many other government and private organizations that have been leading different efforts to raise awareness about the short- and long-term effects of concussions.

A traumatic brain injury (TBI) is defined and indicated by “Any period of loss of or a decreased level of consciousness, Any loss of memory for events immediately before or after the injury, Any alteration in mental state at the time of the injury, Neurological deficits that may or may not be transient, or intracranial lesion following the traumatic event”20. In the United States alone, an estimated 1.7 million TBIs occur each year, leading to more than 1.3 million emergency room visits, a quarter million hospitalizations, and 52 thousand deaths21. The leading causes of TBIs are falls, physical assault/injury, and motor vehicle accidents. In the U. S. Military, over 307,000 cases of TBI have been diagnosed since 2000, 80% of which were in a non-deployed setting22.

Patients who have been screened positive with mTBI are at an increased risk of psychological problems that can have a significant impact in the recovery time. Early detection of psychological conditions such as PTSD following a concussion might improve the overall outcome of a patient and could potentially reduce the cost associated with
treatment.

Dataset

The original dataset consisted of EHR data for 98,342 mTBI patients. The data was filtered to only include patients with more than thirty days of data and no history of moderate or severe TBI. The resulting subsets of 89,840 patients had 5.3 million TBI-related clinical encounters and 8.7 million clinical diagnoses. In this study, a TBI-related encounter was defined as a visit to the doctor regardless of inpatient/outpatient where the patient is treated with one or more of the conditions that are commonly known to be related to concussions such as behavioral disorder, sleep problems, cognitive deficiencies, and audiology complaints. Note that only TBI-related encounters were taken into consideration. The patients under consideration had an average of 59 encounters.

To build our model the dataset was defined to be $P = \{P_1, P_2, \ldots, P_n\}$ where $P$ is a set containing each patient $P_i$ had an associated sequence of encounters $E_i \in \{E_1, E_2, \ldots, E_m\}$ representing unique clinical appointments or hospital visits. Each encounter $E_i$ had an associated set of diagnoses represented by $D \in \{D_1, D_2, \ldots, D_k\}$. To build the sparse matrix representation, as was described in Section , for clustering we set the splitting diagnosis as the first mTBI event and the timeframe of the matrix to be $T = \langle(-60, -30], (-30, 0], [0, 30)\rangle$, where $(-30, 0]$ represents $t_{-30} \rightarrow t_{-1}$ (i.e. 30 days prior to a concussion not including the day of the concussion). The reason for choosing this timeframe was that in our previous work we found that 30 day intervals best capture the critical information of patients and that the first thirty days post concussion are crucial in differentiating between patients.

Figure 9: Example of an automaton, before running through the modified K-Reversible algorithm built from 20 patients that had a concussion (f) and a diagnosis of headaches (H), sleep (S), and/or PTSD (P) with more than 15 days between. For our results, we will be analyzing the trajectory that patients took to develop diagnoses of headaches, sleep, and/or PTSD post concussion, with a minimum of at least 15 days between the concussion and diagnoses being required. We will only be using patients that have enough data to cover the time intervals defined previously, which amounts to 6,473 patients. For generating the automata, we will only be using the first 5,000 patients as any number larger than that requires substantial processing power which is beyond the scope of this paper.

Results

With our dataset defined, we set out to apply the approach that we previously defined in order to construct and analyze a model of the clinical trajectories that patients undergo post concussion to headaches, sleep, and PTSD.

First, we built an automaton consisting of 5,000 patients and ran it against our modified K-Reversible algorithm to identify the usefulness of our algorithm. Figure 9 shows an example of an automaton of 20 patients before being input to our algorithm. In this Figure it can be seen that understanding the similarities and differences between patients is near impossible and this difficulty would grow even more as the number of patients would increase. In Figure 10 we see the computed trajectory for all patients that experience a concussion and end up with a diagnosis of either headache, sleep, and/or PTSD. Analyzing this computed automaton, Figure 10, compared to the initial automaton, Figure 9, reveals that it is much easier to understand the path that most patients take. It should be noted that for the automaton in Figure 10, as with the rest of the computed automata that we will show, we set the line thickness of each transition to be equivalent to the probability of getting the specified diagnosis from that state. These varying thicknesses allow for a viewer to understand the flow of patients visually.

Next, breaking down the trajectories for each individual diagnosis we see headaches in Figure 12a, sleep in Figure
Figure 10: Automaton of all patients developing a diagnosis of headaches (H), sleep (S), and/or PTSD (P) post concussion (f).

Figure 11: Automata of all patients post concussion developing a diagnosis of (a) headaches, (b) sleep, and (c) PTSD.

To understand how the patients can be broken down we input our patients into the k-means clustering algorithm, as defined in the Clustering section of the Approach, to identify the various groups of patients and ran it for $2 \leq k \leq 30$. Using $k$ versus the BIC and the elbow method we found a value of 5 to be the most optimal for splitting the patients into clusters. With this value of 5 for $k$ the number of patients in each group varied, with the counts being: 2543, 95, 820, 3002, and 138 respectively. With the patients clustered into 5 groups, we now would be able to analyze the automata for each group and identify the various trajectories taken.

Using these defined patient clusters, we analyzed the five groups for patients that experienced a diagnosis of headache, sleep, and/or PTSD post concussion. We see the four of the five groups in Figure 12.

In these figures we are able to see the various paths that patients took to these diagnoses and from this we are able to characterize the groups as such:

- **Group 1** Early diagnosis, but long history of encounters afterwards.
- **Group 2** Early diagnosis, encounters, followed by late diagnoses.
- **Group 3** Early diagnoses, followed by a long history of encounters (similar to Group 1).
Figure 12: The first four groups of patients (a, b, c, d respectively) developing a diagnosis of headaches (H), sleep (S), and/or PTSD (P) post concussion (f).

- **Group 4** Early diagnoses and a short length of encounters with high probabilities.
- **Group 5** Early diagnosis, breaks, followed by late diagnoses with a short length of encounters.

With these five groups of encounters we are able to see that we can identify the varying, but most common trajectories that patients take in their diagnoses of headaches, sleep, and/or PTSD post concussion. This information could be passed along to clinicians in the TBI discipline so that they could gain insights that they could not otherwise discover.

**Conclusion**

In this paper we have defined our approach that enables a clinical dataset of patient encounters to be clustered into groups of similarity, represented as NFA’s and DFA’s that include a timeline indicating the number of days between each encounter, and finally run through our modified K-Reversible algorithm to produce automata that display the most common trajectories taken by patients. In our results we applied our approach to an extensive clinical dataset of 89,840 patients that were diagnosed with headache, sleep, and/or PTSD post concussion and showed that with our clustering and grammar induction algorithms we are able to produce trajectories that clearly show the path that each cluster of patients take. With this information we were able to identify and characterize the five different groups of patients that existed in our dataset and with this information we have shown that our approach is effective at clustering and visually representing the clinical trajectory of patients. We anticipate that the automata that we have generated will assist clinicians in understanding the path that their patients take and then being able to provide early treatment to help patients avoid going down specific paths.

**References**


[22] DoD Worldwide Numbers for TBI.
Hierarchical Bayesian Logistic Regression to forecast metabolic control in type 2 DM patients

Arianna Dagliati, Msc¹, Alberto Malovini, PhD², Pasquale Decata, MD², Giulia Cogni, MD²¹, Marsida Teliti, MD², Lucia Sacchi, PhD¹, Carlo Cerra, MD³; Luca Chiovato, MD, PhD²¹, Riccardo Bellazzi, PhD¹²

¹University of Pavia, Pavia, Italy; ²IRCCS Fondazione S. Maugeri, Pavia, Italy; ³ATS Pavia, Italy

Abstract

In this work we present our efforts in building a model able to forecast patients' changes in clinical conditions when repeated measurements are available. In this case the available risk calculators are typically not applicable. We propose a Hierarchical Bayesian Logistic Regression model, which allows taking into account individual and population variability in model parameters estimate. The model is used to predict metabolic control and its variation in type 2 diabetes mellitus. In particular we have analyzed a population of more than 1000 Italian type 2 diabetic patients, collected within the European project Mosaic. The results obtained in terms of Matthews Correlation Coefficient are significantly better than the ones gathered with standard logistic regression model, based on data pooling.

Introduction

Predictive models may play an important role in guiding interventions during monitoring of chronic diseases. However, such models, in particular risk calculators, are often learned from data sets coming from studies based on specific experimental designs and on patients with specific population characteristics that do not (fully) apply to the clinical group of interest. Moreover, the availability of monitoring measurements is typically not considered when building models, as well as the presence of important dynamic factors, such as delivery of drugs and/or lifestyle changes. As a result, the clinical utility of many risk calculators is strongly limited by model calibration drift (due to the different background characteristics of patients and to the different treatment approaches) and by model inadequacy (due to the lack of dynamic information).

In this paper we present an approach to learn predictive models able to take into account both the characteristics of the specific population at hand and the presence of repeated monitoring measurements about the same patients, including changing clinical conditions. In particular we propose to use Hierarchical Bayesian Logistic Regression models, which allow learning individual models borrowing strengths from data coming from all patients thanks to a suitable population model. Our approach is applied to a relevant clinical and social problem, the prediction of metabolic control in Type 2 DM (T2DM) patients; models are learned on the basis of a 10 years retrospective data collection gathered within the EU project MOSAIC.

A related approach can be found in ¹: they tested how fast HbA1c (Glycated Hemoglobin) changes after a change in glucose-lowering medication¹ within a 2-week cohort study. The described analysis has implications for whether routine HbA1c testing intervals before 12 weeks could inform diabetes medication adjustments. In ² authors evaluated the presence of factors associated with HbA1c control over 4 years in an observational cohort study of T2DM patients (CREDIT study) starting insulin therapy. They found that poor glycemic control over the 4 years was mainly modulated by the HbA1c values measured before starting therapy after adjusting for informative covariates (BMI – Body Mass Index, age, diabetes duration, glucose-lowering drugs, basal insulin alone, higher insulin dose and female sex).

Background

Diabetes and the MOSAIC project

Diabetes Mellitus (DM) is a chronic metabolic disorder which encompasses a range of metabolic diseases that share the phenotype of hyperglycemia over a prolonged time¹. Type 2 Diabetes Mellitus (T2DM), also referred to as non-insulin dependent diabetes or adult-onset diabetes, is the most common form of DM, encompassing at least 90% of all cases ². It refers to a heterogeneous group of disorders characterized by variable degrees of insulin resistance (reduced tissue responses to insulin) and insulin deficiency (inadequate insulin secretion for glucose load) ³. T2DM results from the interaction between a genetic predisposition and behavioral and environmental risk factors ⁴. The risk of developing T2DM increases with age, obesity, and lack of physical activity. It occurs more frequently in
individuals with hypertension or dyslipidemia and its frequency varies in different ethnic subgroups. It is often associated with strong familial, likely genetic, predisposition; however, the genetics of this form of diabetes are complex and not clearly defined.

The prevalence of DM in the worldwide population has risen dramatically over the past 2 decades, with 382 million patients diagnosed with diabetes in 2013. Based on current trends, the International Diabetes Federation estimates that by 2030 there will be about 550 million people suffering from DM, making it a leading cause of morbidity and mortality.

The metabolic dysregulation associated with DM causes secondary pathophysiologic alterations in multiple organ systems. There is strong evidence that chronic hyperglycemia plays a causative role in the pathogenesis of diabetic vascular complications. However, the mechanism by which it leads to such diverse cellular and organ dysfunction is still unknown, though several theories have been proposed. The association between increased levels of glycated hemoglobin (Hba1c), which is the reference parameter to measure average plasma glucose concentration over prolonged periods of time, and a higher risk of developing complications has been shown by landmark studies such as the Diabetes Control and Complications Trial (DCCT) in Type One DM patients and the United Kingdom Prospective Diabetes Study (UKPDS) and the Kumamoto study in T2DM patients. The findings of these studies emphasized the importance of intensive glycemic control in all forms of DM and of early diagnosis and strict blood pressure control in T2DM to prevent the adverse effects of the complications of diabetes. A recent study carried on the Italian population suggests that Hba1c variability affects nephropathy more than its average value, thus highlighting the importance of stability of metabolic control in the prevention of microvascular complication.

These recent findings made clear that it is important to retrieve more evidence from longitudinal studies, and to estimate how the variations in glycemic control are affected by heterogeneous factors.

MOSAIC (Models and simulation techniques for discovering diabetes influence factors) is an EU-funded project carried out within the 7th Framework Program, including partners from Italy, Spain, Finland and Greece. The project is devoted to: i) the development of mathematical models and algorithms that can enhance the current tools and standards for the diagnosis of T2DM. Impaired Glucose Tolerance (IGT) and Impaired Fasting Glucose (IFG); ii) improve the characterization of patients suffering from those metabolic disorders and iii) help in evaluating the risk of developing T2DM and its related complications. The final aim of the project is to deploy tools that may contribute to improve the control of the disease evolution, including the prevention of complications, by defining suitable clinical and lifestyle interventions.

One of the main tasks of MOSAIC is to exploit data mining techniques to better understand the mechanisms underlying the progression of diabetes through the analysis of individual patient histories, temporal events and behavioral factors.

Multiple clinical databases at the European level have been made available to the MOSAIC consortium as a result of the activities carried on by its members in previous projects and studies. These data have been exploited for the development of prediction models incorporating information related to environmental and clinical factors including physical, metabolic, phenotypic and lifestyle variables. The objectives of these models are the identification of relevant clinical pathways in patients’ histories and the stratification of the population for the risk of developing T2DM and its related complications. The aim is to integrate them in current decision support systems to enhance decision making in clinical care.

The dataset collected in the Italian Pavia area within the MOSAIC project includes data of 1020 T2DM patients and retrieved from two sources: the hospital Fondazione Salvatore Maugeri (FSM), which collects the clinical data related to the routine clinical activities, and the local public health agency (Agenzia Sanitaria Locale, ASL), which collects process data for administrative and organizational purposes. The integration of these two data sources provides a complete view of the clinical histories of the diabetic patients enrolled in the study. Variables considered in the analysis include demographic data (age, gender, diabetes duration), clinical data from the integrated FSM and ASL dataset (body mass index, glycated hemoglobin, smoking habit, lipid profile and systolic blood pressure) and administrative data collected by the ASL (drug purchases). Durations of the follow-ups of the patients included in our dataset are shown in Figure 1 and range from 1 to 18 years, with a mode of 5 years.

Risk calculators

Within the MOSAIC project, we initially studied the problem of selecting suitable calculators of the risk of complications to be used in our patient group. A score for cardiovascular risk was available, the validated Progetto
Cuore score\textsuperscript{16}. Such score, derived from the Framingham study\textsuperscript{17}, has been adapted to the Italian population. However, FSM treats patients with metabolic control instability and patients suffering from cardiac diseases. For this reason, in the FSM Hospital dataset, microvascular complications have a larger number of cases developed after the first visit as compared to macrovascular ones. Moreover, patients usually start to be treated by FSM (and their data collected) after the disease onset. Given the very nature of the data available in Pavia, we could not find any microvascular risk model that was applicable without a consistent calibration drift.

We therefore decided to build predictive models based on data collected at the first visit for microvascular complications: Retinopathy, Nephropathy and Neuropathy. We considered three prediction horizons: 3, 5, and 7 years from the first visit, and we applied standard modeling strategies, including logistic regression and Naive Bayes. After feature selection and model validation, we selected the models with the best overall performances. As an example, the variables included in the 5 years models were: HbA1c, for all the models; Time from T2DM Diagnosis and Hypertension for Retinopathy; Body Mass Index (BMI), Gender, Hypertension and Smoking habit for Nephropathy; Gender, Age, Time from T2DM Diagnosis, BMI for Neuropathy models.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Follow-up durations.}
\end{figure}

The outline of the above described analyses follows a cross-sectional approach and the risk calculators could be particularly useful when employed during the first visit, to suggest to the clinicians those patients that might need particular attention. They are suitable for the prediction of chronic complications onset. However, these models cannot be used when monitoring patients during their disease progression, especially when the goal is to predict variations in complication risk biomarkers, like Hba1c, during consecutive follow-ups.

As already discussed in previous works\textsuperscript{18–20} another limit in deploying traditional risk calculators for patients monitoring is to cope with external and internal heterogeneity, where external heterogeneity is the one represented by the differences among patients within a population, and internal heterogeneity is the one produced by variations in single patient’s state over time. Various methods\textsuperscript{21,22} have been proposed to overcome these heterogeneity effects, while considering the irregular structure of medical datasets. In the following we will present a Bayesian hierarchical model, able to deal with both kinds of heterogeneity.

\section*{Methods}

\textit{Hierarchical Bayesian Logistic Regression models}

Hierarchical Bayesian Logistic Regression models are of peculiar interest when data are characterized by repeated measures, i.e. follow-up, by units of observation, i.e. patients. In this case two models are jointly applied: one model is used for within-unit analysis, dealing with internal heterogeneity, and another model for across-units analysis, dealing with external heterogeneity. The Bayes theorem is used to integrate the two models and to properly account for the uncertainty in the data, allowing individual learning by borrowing strengths from population data\textsuperscript{23}.
Given $h=1, \ldots, N$ patients, and $i=1, \ldots, n_h$ measurements available on the $h$-th patient, collected on a feature vector $x$ of $m$ “monitoring” variables, and on an outcome binary measure $y$, the probability that the outcome occurs (say metabolic control worsening) is described by the logistic model:

$$P(y_{hi} = 1|x_{hi}) = \frac{\exp (x^t_{hi}\beta_h)}{1 + \exp (x^t_{hi}\beta_h)}$$

The parameter vector of the $h$-th patient, $\beta_h$, is assumed to be a stochastic variable described by a “population” linear model:

$$\beta_h \sim N(\Delta^t z_h, V_{\beta})$$

where $N(\ldots)$ is the Gaussian p.d.f., $z_h$ is a vector of $s$ “static” covariates, such as sex, $\Delta$ is a $s \times m$ matrix of population parameters that associates static and monitoring variables and $V$, is a $m \times m$ covariance matrix.

In order to perform Bayesian inference from the data, the population parameters are typically provided with a suitable prior choice. In our case, we will consider:

$$\text{vec}(\Delta|V_{\beta}) \sim N(\text{vec}(\overline{\Delta}), V_{\beta} \otimes A^{-1})$$

$$V_{\beta} \sim IW(v, V)$$

where vec(.) is the vector representation of the elements of a matrix, $A$ is a suitable prior precision matrix of size $s \times s$, $\otimes$ is the Kronecker product, $IW$ is the Inverse Wishart distribution, and the prior hyperparameters ($\overline{\Delta}, A, v, V$) are typically selected to generate diffuse priors.

The estimate of such models is usually performed by resorting to Markov Chain Monte Carlo (MCMC) methods. Instead of deriving the analytic form of the posterior distribution, the idea behind the MCMC approach is to create a Markov chain able to generate draws from the posterior distribution of the model parameters. These Monte Carlo draws are then used to calculate statistics of interest such as parameter estimates and confidence intervals. Despite the idea behind MCMC methods is simple, its implementation requires the derivations of the appropriate (conditional) distributions in order to produce the draws. Many tools exist to efficiently implement MCMC strategies. In our case, the Hierarchical Bayes logit model is implemented in the R environment (www.r-project.org) package called bayesM24. The bayesM package includes the rhierBinLogit function, which implements an MCMC algorithm for hierarchical binary logits with a normal heterogeneity distribution. This is a hybrid sampler with a Random-Walk Metropolis step for unit-level logit parameters25.

Data pre-processing

HbA1c was originally expressed in mmol/mol and it was converted into % by the following formula: HbA1c % = (0.0915 x HbA1c mmol/mol) + 2.15. Missing values for continuous variables (body mass index, systolic blood pressure, total cholesterol and triglycerides) were imputed by the k-Nearest Neighbor (k-NN) method, using the knnImputation function implemented in the R package called DMwR. Missing values for smoking status were imputed assuming that patients do not change their smoking habit during their follow up period, otherwise by the most frequent value observed in the dataset (for patients for which the smoking status was missing at any visit). Continuous variables were then discretized into three intervals according to the 33th and 66th percentiles of their distribution and recoded as dummy variables. Male gender was recoded as “1” while females as “0”, since the expected risk of diabetes complication is expected to be increased in men. Similarly, smoking status was recoded using dummy variables with “never-smoker” being the category at lower risk, while “ex-smoker” and “current-smoker” at intermediate and high-risk respectively.

Patients were excluded if being characterized by less than 3 follow-up visits. Single visits were excluded if not followed by at least one visits within a time frame of 12 months. Figure 2 shows the distribution of time spans between follow-ups in the population.

The dependent variable of the analysis was defined as an observed increase in terms of HbA1c % $\geq 0.5$ within a time frame of 12 months from each visit. The presence of the increase was coded as “1”, “0” otherwise. Demographical and behavioral characteristics (Gender, Age at follow-up visit, Time between follow-up visit and Time from T2DM diagnosis, Smoking habit), clinical measurements (HbA1c, BMI, Triglycerides, Systolic Blood Pressure (SBP), Total Cholesterol) and presence of pharmacological treatments (Insulin, other Anti-Diabetic treatments) were considered as independent variables to be included in the analysis.

---

1 The output is the covariance block matrix for vec(A) of size $m \times s \times m \times s$. 
Figure 2. Time between follow-up, in months.

Data deriving from the first three visits available for each patient were used to train the model. At the first iteration, predictions were performed on data deriving from the 4th visit and the event to predict was the occurrence of an increase in terms of HbA1c % of at least 0.5 with respect to HbA1c % measured at 4th visit in a time-frame of 12 months. To be noted, only a fraction of the whole cohort has data for more than 3 visits, thus the prediction is performed on a subset of subjects. At the second iteration, measurements deriving from the 4th visit were included in the training set and the prediction was performed on data from the 5th interval and so on until the 6th iteration (prediction on visit number 9). Figure 3 provides a graphical representation of the analysis schema.

Figure 3. Analysis schema. Example of analysis involving three patients with a number of visits ranging from 4 to 5. Blue and red rectangles correspond to visits included in the training set and test set respectively. Grey rectangles
correspond to visits not considered at the specific iteration. At the first iteration data from the first 3 visits were used to train the model that was tested on data deriving from the 4th visit (all 3 subjects had information about HbA1c % for at least one year starting from the 4th visit). At the second iteration, data from the 4th visit was included in the training set: once learned, the model was tested on data from the 5th visit (only patient A and C had information about HbA1c % for at least one year starting from the 5th visit).

The same analysis schema was performed in parallel by logistic regression (glm function with logit link in the R package stats). The deriving regression coefficients were used to estimate the probability of the outcome on unseen data.

The probabilities estimated by the Hierarchical Bayesian model and by logistic regression were binarized using the probability of increase in terms of HbA1c % ≥ 0.5 characterizing the training data. Once the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications were estimated, models performances were quantified and compared in terms of Matthew’s Correlation Coefficient that is the most appropriate metrics to be considered in presence of unbalanced dependent variables.

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}}
\]

**Results**

**Cohort characteristics**

A total number of 684 T2D patients (401 males, 283 females) meeting the inclusion criteria were analyzed, with a median number of 10 visits (min = 3 – max = 52). Table 1 describes the general characteristics of the analyzed cohort at the first and last visits considered.

<table>
<thead>
<tr>
<th>Variable</th>
<th>First visit</th>
<th>Last visit</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking History (Yes/Ex/Never)</td>
<td>113/224/347</td>
<td>113/224/347</td>
<td>\</td>
</tr>
<tr>
<td>Age (years)</td>
<td>62 (56-69)</td>
<td>67 (56-69)</td>
<td>\</td>
</tr>
<tr>
<td>Time from T2DM diagnosis (years)</td>
<td>1.88 (0.64-8.41)</td>
<td>7.91 (0.64-8.41)</td>
<td>\</td>
</tr>
<tr>
<td>BMI (Kg/m²)</td>
<td>29.02 (26.13-32.56)</td>
<td>28.62 (26.13-32.56)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Total Cholesterol (mg/dl)</td>
<td>197.78 (172.14-224)</td>
<td>178.47 (172.14-224)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>SBP (mm Hg)</td>
<td>140 (126.11-150)</td>
<td>130 (126.11-150)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Triglycerides (mg/dl)</td>
<td>136.01 (106-188.35)</td>
<td>122.6 (106-188.35)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>HbA1c %</td>
<td>7.2 (6.5-8.7)</td>
<td>6.8 (6.5-8.7)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Insuline</td>
<td>67 (9.8%)</td>
<td>94 (13.7%)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Other drugs</td>
<td>323 (47.22%)</td>
<td>468 (68.42%)</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of the analyzed patients at first and last visits considered. Distributions are described by counts (frequency, %) or by median (25th – 75th percentiles); p – value = p – value from the Wilcoxon signed rank test or from the McNemar test comparing variables distribution between the two visits.

**Results from the rhierBinLogit function**

The rhierBinLogit function allowed estimating population- and individual-level regression coefficients for each time-varying covariate included in the analysis (age at follow-up visit, time between follow-up visit and T2D diagnosis, HbA1c, BMI, triglycerides, SBP, total cholesterol, insulin, other treatments) conditioned on the static variables considered (smoking habits, gender). Table 2 reports the regression coefficients estimated using data from the first 3 visits.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>Intercept</th>
<th>Gender (M)</th>
<th>Smoke (Yes)</th>
<th>Smoke (Ex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>[61-70)</td>
<td>-0.77</td>
<td>1.74</td>
<td>-2.56</td>
<td>-0.12</td>
</tr>
</tbody>
</table>

475
Table 2. Regression coefficients. Each column represents a static variable, each row a different time-varying variable and corresponding intervals defined by the 33th and 66th percentiles of the distribution (the lowest interval was used reported since it was considered as the baseline).

Individual-level regression coefficients learned on the training sets were used to estimate the probability of experiencing an increase in terms of HbA1c % ≥ 0.5 within a time frame of 12 months from each visit. Table 3 reports the total number of patients and corresponding frequency of events in each training set and test set considered, according to the analysis schema in Figure 3. Regarding the computational complexity of the method, the algorithm took an average time of 14.83 ± 0.14 minutes to fit each model (computer characteristics: iMac, 4GB RAM), for a total computational time of 88.89 minutes for performing the whole analysis (six models were fitted). Logistic regression was also applied, as a term of comparison. Results are reported in Table 3 and show that the MCC reached by Hierarchical Bayesian Model was always higher than MCC obtained by logistic regression on the same data except for the prediction after the 9th visit (t-test p-value <0.05).

Table 3. Dependent variable's distribution on the training and test sets and discriminative performances on the test set. Learned on data from visit # (training set) = visits used to train the model and defining the training set; Prediction on data from visit # (test set) = visit on which the model was tested and defining the test set; Training set: n. patients, mean n. visits = number of patients defining each training set and average number of visits; Test set: n. events/ n. tot patients (%) = number of patients experiencing an increase in terms of HbA1c % ≥ 0.5 within a time frame of 12 months from the visit considered as test set / number of patients observed during the visit considered as test set; MCC Hierarchical Bayesian Model = Matthew’s Correlation coefficient reached by the Hierarchical Naïve Bayesian Model in discriminating patients experiencing an increase in terms of HbA1c % ≥ 0.5 within a time frame of 12 months from the visit considered as test set from those who did not experienced such increase; MCC Logistic Regression.
Logistic Regression Model = Matthew’s Correlation coefficient reached by the Logistic Regression Model in discriminating patients experiencing an increase in terms of HbA1c % ≥ 0.5 within a time frame of 12 months from the visit considered as test set from those who did not experienced such increase.

Conclusion

A challenging problem of clinical data mining is to derive predictive models able to be tailored to both the specific clinical center and to the single patients’ characteristics, i.e. to take into account external and internal heterogeneity. This is crucial when monitoring chronic patients, where repeated measurements are available on a potentially large cohort of subjects. In order to deal with this problem, we have proposed a method based on a Hierarchical Bayesian logistic regression approach, which is able to suitably take into account individual information when the individual model parameters are estimated, while exploiting the data of other patients thanks to a population model. When applied to the problem of predicting variations in HbA1c between two consecutive visits of T2DM patients, the model proved to outperform standard logistic regression modeling based on data pooling. The approach seems therefore promising to deal with this class of problems, which turn out to be very common in the clinical context. Although one of the limitations of the obtained results is related to the choice of a definite population, coming from a single hospital in Italy, where patients show specific characteristics. For this reason, the obtained model may not be generalized well to other patient populations. Moreover, the method itself has two potential drawbacks. First, parameters estimate relies on MCMC methods. This makes the estimation strategy inefficient and not easily scalable when very large data sets are available. Second, the model assumes that the measurements are exchangeable, i.e. time is not explicitly modeled in the individual measurements. In the future we will work on improving both aspects; first in trying different estimation strategies, such as variational methods; second, we will consider autoregressive structures when modeling dynamic features.

Acknowledgment

This work is part of the MOSAIC Fp7 project, funded by the European Union. We want to thank Camilla Colombo and Monica Cigognini for their support in data analysis and models implementation.
References


Design and Evaluation of a Medication Adherence Application with Communication for Seniors in Independent Living Communities

Dipanwita Dasgupta¹, Reid A. Johnson, PhD¹, Beenish Chaudhry, PhD¹, Kimberly G. Reeves, MPA², Patty Willaert, MPA², Nitesh V. Chawla, PhD¹
¹ University of Notre Dame, Notre Dame, IN; ² Memorial Hospital of South Bend, South Bend, IN

Abstract

Medication non-adherence is a pressing concern among seniors, leading to a lower quality of life and higher healthcare costs. While mobile applications provide a viable medium for medication management, their utility can be limited without tackling the specific needs of seniors and facilitating the active involvement of care providers. To address these limitations, we are developing a tablet-based application designed specifically for seniors to track their medications and a web portal for their care providers to track medication adherence. In collaboration with a local Aging in Place program, we conducted a three-month study with sixteen participants from an independent living facility. Our study found that the application helped participants to effectively track their medications and improved their sense of well-being. Our findings highlight the importance of catering to the needs of seniors and of involving care providers in this process, with specific recommendations for the development of future medication management applications.

Introduction

Seniors take on average about seven medications per day, representing the highest number of prescribed medications and the most complex medical regime compared to any age group. While this complexity is a natural response to chronic disease conditions that are increasingly prevalent as one ages, it also brings with it a higher risk of medication non-adherence. Medication non-adherence increases from an average of approximately 20% for patients taking one medication daily to over 50% for those taking medications four times per day. It not only results in unalleviated disease conditions, but can also lead to an increased risk of comorbidities and longer hospital stays. Ultimately, it is responsible for approximately 125,000 deaths per year at an annual cost of roughly $100 billion.

Mobile applications have been proposed as promising tools for mitigating some of the challenges of medication non-adherence by providing seniors with tracking, organizational, and reminder supports. Unfortunately, there are several reasons why such applications continue to have limited impact on the aging population. The most pronounced problem is that the vast majority of these applications are designed for technologically savvy individuals, fundamentally limiting their use by a senior population that is often slow to adopt new technology. Further, despite the integral role that communication between patients and care providers plays in providing healthcare workers with the root causes of non-adherence, these applications generally have very limited ability—and often no ability—to connect the patient with care providers. These factors severely limit the practical use of these applications by seniors.

We are developing eSeniorCare, a comprehensive system targeting quality of life, comprising both a tablet application and a web portal. While the application has an extensive set of functions, in this paper we focus specifically on the application’s medication adherence component. This component assists seniors in tracking their medications by providing reminders for when medication is due and by recording the intake of medications. It also provides a communication mechanism with the care provider. The web portal helps care provider teams to communicate with seniors about missed medications, changes, and refills. By building the system around these components, eSeniorCare has been carefully designed with the features needed to empower seniors to take charge of their medication management and to enable care providers to support the medication adherence of their patients.

To investigate the impact of eSeniorCare on medication adherence, we conducted a pilot study of the system at a local independent living facility catering to a low-income senior population. The care provider team at this facility consists of a nurse and Resident Life Health Administrator (who we collectively refer to as RLHA) under the supervision of Memorial Hospital of South Bend (part of Beacon Health System). We present the design, development, and evaluation of the system’s medication management component to investigate two main research questions (RQ1 and RQ2) and an exploratory research question (RQ3):

• RQ1: How do the participants use the application to track their medications?
• RQ2: How does the application impact participants’ quality of life, specifically as it pertains to health?
• RQ3: What influence does the application have on participants’ knowledge of their medication regimes?

Related Work

There are various methods for measuring medication adherence. Some of the methods measure adherence directly, while others attempt to infer adherence through indirect (proxy) indicators. Direct methods include measurements that determine drug metabolites, biomarkers or additionally applied marker molecules in biological fluids (e.g., plasma and urine), and direct observation therapy. While these methods can provide fairly accurate measurements of adherence, they are not scalable, incur high costs, and tend to be labor intensive. Indirect methods include electronic pill box cap monitoring, pill counts, self-reporting, and pharmacy refill systems. While these indirect methods are fairly inexpensive, their reliability depends on the individual’s responses. For example, an individual can open the electronic pillbox for a medication despite not physically taking it, thereby providing an overestimation of adherence.

The increasing popularity of smartphones has led to the development of mobile applications for medication management. These applications typically provide reminders for when a medication is due, information on medication dosage history, and medication refill alerts. While some of these applications have been developed without a focus on any particular condition, many have been developed to cater to chronic disease conditions such as hypertension, diabetes, and asthma. The potential of these mobile applications to address the challenges of medication management has been encouraging, with several studies demonstrating that they can be effective in improving medication adherence.

Despite their potential, very few mobile applications have been developed for or targeted at seniors. One such application, ALICE, provides medication reminders and can send messages to the care providers in the case of non-response to these reminders. However, almost none of these applications track the medications taken as needed (PRN) and many provide no or non-real-time information regarding missed medications. Further, few of these applications support a direct connection between the patient and care provider. As a result, unlike eSeniorCare, these applications do not enable intervention by the care provider in the case of medication non-adherence. We have also designed eSeniorCare with several features practically useful for medication management, such as care provider follow-up in case of missed medications and PRN intake, that are missing from currently available mobile applications.

Application Design

We are developing a tablet-based Android application called eSeniorCare that focuses specifically on the components of quality of life for seniors. In this section, we provide an overview of the features provided by the application’s medication management component, as illustrated in Figure 1(a).

Medication Reminders. For each participant, eSeniorCare automatically generates reminders for all prescribed medications based on the prescription data that is queried from the encrypted local database (Figure 1(b)). A reminder is scheduled five minutes before the prescribed dosage time. For each medication due at the dosage time, the reminder screen displays the following information: (a) medication name, (b) dosage information (i.e., quantity and unit), (c) associated treatment condition, (d) audio icon to speak out the medication’s name, (e) image of the medication, and (f) a checkbox that the user must check to confirm intake. If the checkbox associated with any medication is not checked, the user is asked to validate the response. The user can also view a magnified version of the image of the pill.

Medication History. For each participant, all current and previously prescribed medications are listed, along with a status—active, dosage modified, or discontinued—for each medication. This feature can help participants improve communication with their physicians.

Current Medications. The application provides participants with comprehensive information about their current medications, along with the frequency and dosage of each medication and the corresponding treatment condition. The purpose of this feature is to empower the participants in maintaining an aggregated list of prescribed medications, which may be used as a quick reference for the patient or clinician during medical visits.

PRN. Pro re nata (PRN) medications are taken whenever the need arises. All PRNs are listed here, along with corresponding disease diagnoses, and dosage directions. Due to their potential impact on patient health, the inclusion of PRN medications is an important feature that has traditionally been absent from mobile health applications.
Medication Taken. The participants may respond to a medication reminder even after it has disappeared by tapping the “Taken” button, as shown in Figure 1(b). The participant is then presented with a list of dosage times before the current system time for which responses have not yet been entered. If a participant selects a particular dosage time, he or she is presented with the appropriate medication reminder. This feature is typically unavailable in the mobile health applications.

Medication Change. A participant can inform the RLHA to a change in his or her medication by clicking this button. The RLHA receives information on the medication change via the web portal. The RLHA then follows up on this information by updating the participant’s medication information.

RLHA Portal for Monitoring Adherence. A participant’s response or non-response to the reminders is reported to the RLHAs via the web portal, allowing the RLHA to follow up with the participant to resolve any issues and to provide appropriate motivation. To validate the participant responses for medication reminders, we implemented the “Refill Medications” button that provides a list of the refillable medications. They can pick the ones for which they require a refill. We calculate an estimate of the refill date for each refillable medication based on the initial pill count and the dosage. If the requested date does not match the estimated refill date, the RLHA is notified through the portal.

Missed medication is marked “Not Taken”; otherwise, it is marked “Taken”. In order to track the workflow of a missed medication and for ease of maintenance, we added a status field. When no action has been taken by the RLHA, the status is “Unread”. After the RLHA has followed up with the participant, the status is changed to “Resolved” and a reason is entered for the missed medication. The RLHA may use the portal to create new users, populate and maintain their medication profiles.

Offline Usage. Our study site has Wi-Fi connectivity that is limited to the core of the building and community areas. As a result, the majority of the participants did not have Wi-Fi connectivity in their apartments. To account for the intermittent connectivity, the tablet application has a local database—located in the tablet’s encrypted internal memory—that enables the desired offline functioning of the application. The application is only required to connect to Wi-Fi once a day for synchronization with the remote server.

Methods

We conducted a study of eSeniorCare in collaboration with the Aging in Place (AiP) program designed by the Community Health Enhancement (CHE) vision of the Memorial Hospital of South Bend. The program operates in several local independent living facilities with a mixed-income senior population, and aims to positively impact wellness and quality of life for its participants while they are living in their current home settings. As part of this aim, AiP is actively interested in the development of a low-cost solution to help its participants improve adherence to their medication regimens.
Table 1. Medication Adherence Questionnaire (MAQ) provided to participants.

<table>
<thead>
<tr>
<th>Description</th>
<th>Options</th>
</tr>
</thead>
</table>
| Match the diseases diagnosed and medications taken. | • List of diagnoses  
• List of medications |
| Fill out the following details [options] for 2 medications. | • Diagnosis  
• Dosage information (quantity per intake and number of times per day)  
• Identification (color, name, size, pill bottle / container for daily use) |
| How many pills do you take per day? | No options provided |
| How do you remind yourself of the medications to be taken? | No options provided |

Overview: We held an interest meeting with residents at the facility to introduce the purpose of the study and requirements of participation. We only accepted residents who were cognitively intact, determined by the St. Louis University Mental Status Exam (SLUMS), a 30-point screening questionnaire that tests for orientation, memory, attention, and executive functions; fluent in English; and had good vision. Finally, participation was restricted to individuals with at least one chronic condition and one actively prescribed medication.

Each resident who qualified for the study, based on the selection criteria described above, was given an encrypted, password-protected 7-inch Samsung Galaxy Tab 2 tablet device. Participants were expected to carry the device with them at all times and to manage their medications as reasonably directed by the device. Those who participated for the full duration of the study were permitted to keep the tablet.

Participants. We recruited sixteen participants with an average age of 66 years (SD = 9.2). Five participants were male and eleven were female. Twelve participants identified themselves as African American/Black, three as non-Hispanic White, and one as other. One participant was married, six were divorced, two were widowed, two were separated, and five never married. One participant was employed, one was looking for work, seven were retired, and four indicated that they could not work. Two participants had between 9th and 11th grade education, four had completed high school, eight had 1–3 years of technical school, one was a college graduate, and one had an advanced degree. The number of medications per participant ranged from 1 to 13, with an average of 6–7 medications.

Study Design. We conducted the study over a period three months to evaluate the RQ1, RQ2, and RQ3. The participants were required to respond to the medication reminders, report PRN intake, and report changes in medication. To familiarize participants with the application, one-on-one training with the application was provided at the beginning of the study. The RLHA were instructed to access the web portal on alternate days to track the medication responses. In the case of missed medication, a registered nurse followed up with the participant and took appropriate action.

Evaluation Metrics. To answer our research questions, we used the following information:

- Medication Usage Logs (MUL): Participant responses to medication reminders, reported changes to prescribed medication, and incidences of PRN were recorded in a database.
- Medication Adherence Questionnaire (MAQ): A questionnaire designed based on the Brief Medical Questionnaire (BMQ) for measuring medication adherence. Studies have shown that if a patient is better aware of her or his medications and diseases, then he or she will have an improved adherence to his or her medication regimen. We administered a paper-based MAQ both at the beginning and at the end of the study. Table 1 provides an outline of the questionnaire. The questions in MAQ are tailored to the prescription information provided by the participants. The response(s) for each question are matched with the prescription information.
- Final Feedback Questionnaire (FFQ): A 38-point post-study feedback questionnaire for evaluating the usability of the application and the effectiveness of the study. Sample questions are included in Table 2. The questionnaire was administered both on paper and online at the end of the study.
- SF12v2: A 12-item health survey provided by Quality Metric to measure the general health (physical and mental) of an individual and health-related quality of life, and well-validated for the aging population. It measures the health status in eight functional domains: bodily pain, physical function, role physical, role emotional, mental health, vitality, social function, and general health. These domains are further collapsed into a physical component summary (PCS) and a mental component summary (MCS). The survey was administered on paper at the beginning and end of the study.
Table 2. Questions and response trends for the Final Feedback Questionnaire.

<table>
<thead>
<tr>
<th>Num.</th>
<th>Question Text</th>
<th>Feedback Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>eSeniorCare made me more aware of my health.</td>
<td>3 6 5 1 1</td>
</tr>
<tr>
<td>Q2</td>
<td>I felt an improvement in my health/well-being as a result of eSeniorCare.</td>
<td>1 7 6 1 1</td>
</tr>
<tr>
<td>Q3</td>
<td>eSeniorCare helped me to know my medications: dosage, purpose for medication etc.</td>
<td>2 5 6 1 2</td>
</tr>
<tr>
<td>Q4</td>
<td>Overall, eSeniorCare helped me manage my medications.</td>
<td>1 9 3 2 1</td>
</tr>
<tr>
<td>Q5</td>
<td>I showed eSeniorCare to my physician.</td>
<td>7 2 5 2 2</td>
</tr>
<tr>
<td>Q6</td>
<td>eSeniorCare showed my correct medication information.</td>
<td>1 13 1 1</td>
</tr>
<tr>
<td>Q7</td>
<td>I would like to enter my own medications.</td>
<td>6 10 1 1</td>
</tr>
</tbody>
</table>

Note: The bar segments from left to right correspond to the responses “Strongly Agree” (green), “Agree” (orange), “Neither Disagree Nor Agree” (blue), “Disagree” (purple), and “Strongly Disagree” (brown). For each segment, the value denotes the number of participants with the response, while its length denotes the proportion.

Results

In this section, we provide details of the results from our study of eSeniorCare. Our results come from information gathered by tracking how participants used the application over the course of the study, as well as from questionnaires, administered at the beginning of the study and at its conclusion.

RQ1: How do participants use eSeniorCare to track their medications?

A majority of the participants \((n = 12)\) initially faced difficulties in maneuvering the touch screen (sensitivity, pointing, and dragging) and with the tablet itself due to a lack of technological fluency and unfamiliarity with the tablet. Several hour-long group workshops and one-on-one training sessions helped familiarize participants with the technology and application. To help overcome usability challenges with the touchscreen, we provided each participant with a stylus. We also provided printed manuals for eSeniorCare application. These activities were substantial in helping to foster user engagement, with participants expressing a strong desire for the study to continue as it approached its conclusion.

The FFQ included several questions to assess how participants used the application to track their medications. All but one of the participants reported that the application showed the correct medications (Table 2, Q6). We note that one participant also made a considerable number of changes to his or her medication list, so it could be a function of that artifact. Around 50% of the participants reported that they showed the application to their physician (Table 2, Q5). More than 65% of the participants reported that they felt the application helped them to manage their medications (Table 2, Q4), and about 50% of the participants reported that they felt the application made them more aware of their health (Table 2, Q1).

Medication Reminders. We used the MUL to define a rating system for tracking the frequency of medication responses for each participant. A rating of 1 was given if the participant acknowledged a medication reminder. The maximum possible daily rating for the participant is then equal to the number of times per day that a medication is due to be taken. We calculated the weekly rating as the sum of the participant’s daily ratings for a week divided by the sum of the participant’s maximum possible daily rating for a week (thus taking into account the variation in the number of daily reminders among the participants). We averaged the weekly rating to obtain the average weekly rating (AWR).

To identify similar usage patterns among participants, we performed \(k\)-medoids clustering on the participants’
weekly ratings, using the measure of silhouette width to evaluate the number of clusters. Based on the maximum average silhouette width, we discovered three distinguishable clusters of participants, illustrated in Figure 2(a). The majority of the participants (n = 7) had moderate usage, shown in Figure 2(b), with AWR ranging from 0.52 to 0.73. The remaining participants were split between low-usage ratings (n = 5), shown in Figure 2(c), with AWR less than 0.28, and high-usage ratings (n = 4), shown in Figure 2(d) with AWR greater than 0.93.

To determine whether the ratings (usage patterns) given by participants grouped into different clusters is statistically significantly different, we performed a one-way ANOVA test on the AWR for the clusters. The extremely low p-value for the AWR of different clusters (Table 3: AWR) indicates that the ratings are statistically significant different. There is also a statistically significant difference in the AWR for different clusters when all of the demographic factors (gender, race, education, employment status, and marital status) are considered as variables in the model (Table 3: DWR). To identify differences between specific groups, we use the post-hoc Tukey-Kramer method (Table 3: AWR right side), which shows that there is a statistically significant difference in the AWR between several pairs of groups.

Table 3. One-way ANOVA results for participants clustered according to response type.

<table>
<thead>
<tr>
<th>Type</th>
<th>F-test†</th>
<th>p-value</th>
<th>Low–Medium</th>
<th>Medium–High</th>
<th>High–Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWR</td>
<td>101.730 (2)</td>
<td>1.15 × 10^-8*</td>
<td>5.15 (4.7 × 10^-6)</td>
<td>5.17 (4.5 × 10^-6)</td>
<td>10.32 (0.00)</td>
</tr>
<tr>
<td>DWR</td>
<td>79.179 (2)</td>
<td>5.35 × 10^-6*</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MAQ</td>
<td>0.084 (1)</td>
<td>0.777</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PCS</td>
<td>6.647 (1)</td>
<td>0.0257*</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MCS</td>
<td>11.381 (1)</td>
<td>0.00621*</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PCSWR</td>
<td>0.329 (1)</td>
<td>0.5778</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MCSWR</td>
<td>1.872 (1)</td>
<td>0.1985</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Note: AWR: Medication Average Weekly Rating; DWR: Medication Average Weekly Rating by Demographics; MAQ: Medication Adherence Questionnaire. Low, Medium, and High refer to the groups shown in Figure 2. * Significant at a 95% confidence level. † Formatted as: score (degrees of freedom). ‡ Formatted as: mean difference (p-value).

RLHA Interaction. Table 4 shows the percentage of medication alerts to which each participant responded, and the number of days over which these responses occurred. A low response percentage was observed for two participants, namely 1 and 13. Participant 1 had only one medication listed as a daily medication, but was advised by his or her physician to take it as needed. Participant 13 had a considerable number of changes to his or her medication list, making it difficult to finalize the list. Table 4 shows the count of medications that were reported “Taken” and “Not Taken” for each participant over the study span. The most common reason cited for missed medications was “did not want to take” (28%), followed by “took but didn’t enter” (17%) and “discontinued” (15%). The RLHA followed up with the participants in case of any missed medications, and medication change. Only 37.5% of the participants would prefer to enter their own medications (Table 2: Q7). All of the participants appreciated this follow-up and felt that their best interest is being looked after.

PRN Intake. We found that eight participants had at least one PRN medication listed during the study phase. For those with at least one PRN medication, the number of PRN medications ranged from 1 to 5, with a median of 5. We also found that five participants tracked their PRN medications at least once over the study period; two of these participants tracked their PRN medications exactly 1 time, two tracked 5 times, and one tracked 69 times. Of the participants who did not track even once during the study phase, one had only a single medication listed and did not receive any confirmation from the physician regarding medication intake.

RQ2: How does eSeniorCare impact participants’ health-related quality of life?

SF12v2 was completed by nine participants. The majority of these participants belonged to the medium-usage group, as defined above. We used the missing data estimation supplied by the SF12v2 scoring software to estimate scores for the remaining participants. A majority of the participants (56.25%) had a post-study physical health score (PCS) score greater than or equal to that at the beginning of the study. We also observed a decrease in mental health score (MCS) for the same proportion of participants at the end of the study. We note, however, that about 50% of the
Figure 2. Cluster analysis based on participants’ weekly ratings. (a) Illustrates the average silhouette width (Y-axis) as a function of the number of clusters (X-axis); the blue point (n = 3) represents the natural number of clusters determined by the silhouette width. Based on the usage patterns, we call these clusters: low-, medium-, and high-usage clusters. (b)(c)(d) Illustrate the weekly ratings for participants of the low-, medium-, and high-usage clusters, respectively; for each plot, the X-axis is the week and the Y-axis is the weekly rating for each of the cluster’s participants, with each line representing a participant.

Table 4. A summary of the frequency, duration, and content of participant responses to medication alerts.

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Number of days</th>
<th>Response Percent *</th>
<th>Taken</th>
<th>Not Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>8.99</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>83.15</td>
<td>819</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>19.10</td>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>73</td>
<td>82.02</td>
<td>186</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>74</td>
<td>83.15</td>
<td>672</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>79</td>
<td>88.76</td>
<td>581</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
<td>65.17</td>
<td>121</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>20.22</td>
<td>94</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>78</td>
<td>87.64</td>
<td>532</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>76</td>
<td>85.39</td>
<td>168</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>83</td>
<td>93.26</td>
<td>436</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>81</td>
<td>91.01</td>
<td>513</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>4.49</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>12.36</td>
<td>66</td>
<td>14</td>
</tr>
<tr>
<td>17</td>
<td>52</td>
<td>58.43</td>
<td>138</td>
<td>94</td>
</tr>
<tr>
<td>19</td>
<td>62</td>
<td>69.66</td>
<td>61</td>
<td>2</td>
</tr>
</tbody>
</table>

* Calculated based on the number of days in the second phase (89 days).

participants reported PCS higher than the U.S. norm for their age category, both at the beginning and end of the study. To examine the relationship between post-study scores and medication component usage (WR), we developed a linear regression model with the post-study SF12v2 component score as the dependent variable and the pre-study SF12v2 scores and cluster number (high, medium and low rating groups) as independent variables, while controlling for demographic variables (age, education and race), other components of the application, and tablet use. We observed a strong relationship between the post-study and pre-study scores (Table 3: PCS and MCS). However, we did not observe a notable relationship between post-study scores and WR (Table 3: PCSWR and MCSWR). This may be due to the short duration of the study, the weather conditions (winter), and the small sample size. Further, we did not take into account the impact of chronic conditions on the scores.

RQ3: What influence does eSeniorCare have on participants’ knowledge of their medication regimes?

To explore the influence of eSeniorCare on a participant’s knowledge of their medications, we examined each participant’s responses to the questions in the MAQ. For each correct response to a question, we gave the participant 1 point. We then calculated each participant’s MAQ score as the total number of points he or she earned divided by the maximum possible points for that participant.
There was an increase in MAQ scores for 50% of the participants, while 19% showed a decrease. The remaining 31% of the participants showed no change in their scores because they obtained maximum or nearly maximum scores for both pre- and post-study questionnaires. To determine whether the difference between the pre- and post-study MAQ scores is statistical significantly different, we performed a paired t-test ($t$-statistic: $-1.03; p$-value: 0.317). We did not find any statistically significant difference. We further examined the relationship between MAQ scores and the clusters assigned based on weekly ratings via a one-way ANOVA test (Table 3, MAQ). We did not observe any statistically significant relationship between the change in MAQ scores and weekly ratings.

Discussion

The main objective of our study was to evaluate the design and impact of the medication adherence component of eSeniorCare on the participant’s health-related quality of life and explore the impact on their knowledge of medications. As a part of three-month-long study, the participants (seniors aged 55 years or older) used a tablet-based application to record their medication intake. The RLHA used a web-portal to track missed medications, changes in medications, and refill alerts reported by the participants through the application. All of the participants completed the study.

From our clustering illustrated in Figure 2 we found that the majority of participants ($n = 11$) used the medication component to track their medications, with most of the participants ($n = 13$) reporting higher (or maximum) MAQ scores at the end of the study than at the beginning. The majority of participants ($n = 10$) also reported higher physical health SF12v2 scores. However we found that, after controlling for demographic factors, the medication component usage (technology) did not have an impact on these scores. This accords with prior studies that have investigated the impact of computer use on quality of life. That said, we did observe a significant change in the post-study scores compared to the pre-study scores, even after controlling for age, gender, education and race.

From our ANOVA tests shown in Table 3 we found that the majority of the participants had an increase in MAQ scores at the end of the study, though the improvement in the MAQ scores for the participants at the end of the study was not statistically significant. We note, however, that the majority of the participants already had almost perfect scores at the beginning of the study, resulting in a very small margin of potential improvement for most of the participants.

For tracking medication adherence, we relied on each participant’s self-reported responses. As there is no gold standard in tracking adherence, a combination of methods is instead recommended. We thus implemented a secondary method of tracking medication adherence via a “Refill Medications” function. During the study phase, however, most of the participants did not use this functionality. This may be because most of the participants had already obtained 90-day refills for their medications, thus obviating the need to use the refill functionality throughout the duration of the study.

Recommendations for future work. We believe that our study provides a foundation for further development of integrated technological solutions for seniors. We also believe that the benefits of these solutions can and should go beyond the seniors themselves, as they can also help care providers and clinicians by providing them with a direct connection for monitoring medication adherence and connecting with their patient pool on health and wellness. Based on our findings, we provide the following recommendations for the development of mobile applications for medication management, and particularly for those oriented to low-income aging populations:

- The application should function appropriately with intermittent Internet connectivity. We found that limited access to Wi-Fi was a major obstacle for the deployment of our application. To use the available connectivity efficiently, we specifically recommend periodically synchronizing locally stored data with a remote server. The care provider should be able to track the last synchronization date for each participant for tracking medication adherence to overcome the challenges with limited to no internet connectivity in a low-income, independent living facility.
- The application should be accompanied by a companion web portal, accessible to a care provider. The portal should provide information on missed medications, refill alerts, and medication changes so that the care provider can take appropriate action. We found that participants were grateful for follow-ups by their care provider team.
- The companion web portal should have functionality for the creation and maintenance of medication profiles for the users. We found that the complex medical regimen maintained by many seniors frequently makes medication entry directly via the application difficult. Further, to expedite this process, the application could be integrated with health information exchanges (HIEs) with necessary safety and privacy measures.
- The participants should be able to send to and receive messages from RLHA using the application as we found this communication to be a key factor for using eSeniorCare.
**Limitations.** While our current study was informative in impact as well as needs assessment, we do believe that having a larger sample pool over a longer time period can help strengthen the findings and conclusions drawn from the study. However, this is an important first step in developing an application that is integrative of the senior participants and their care provider team, demonstrating its strengths and weaknesses, and laying the foundation for partnering with independent living facilities for longer-duration studies. A further limitation is that we only address unintentional medication non-adherence, as our medication tracking relied on self-reporting; this is an area of ongoing research. It should be further noted that other features of the application aside from those actively investigated may have also contributed to the positive findings of our study.

**Conclusion**

We presented *eSeniorCare*, a health and wellness system focused specifically on improving senior care and quality of life. By analyzing participant information collected both through this system and through surveys, we found that *eSeniorCare* provides an intersection of technology and relationship building that can help seniors to effectively track their medications. We observed that the participants became more aware of their medications and were more likely to use them as prescribed. We also discovered that the role of RLHAs—providing personal support and medication tracking—is crucial to this improvement. Based on our findings, we also provide a series of suggestions for future developers to improve the design of their own medical management applications. Through an easy-to-use application and web portal, *eSeniorCare* empowers seniors to take charge of their medication management and places them in the center of a system of support, representing an important step toward improving senior care and quality of life.

**References**

36. Maruhash M. User’s manual for the SF-12v2 Health Survey; 2012.
Investigating Delays in Updates to Infusion Pump Drug Limit Libraries

Poching C. DeLaurentis, Ph.D.1, Kang-Yu Hsu, M.S.1,2, Ana Isabel De la Hoz Armenta1,3, Yuval Bitan, Ph.D.4
1Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, IN; 2School of Industrial Engineering, Purdue University, West Lafayette, IN; 3Universidad Nacional de Colombia, Medellin, Colombia; 4Department of Industrial Engineering and Management, Ben Gurion University of the Negev, Be' er Sheva, Israel

Abstract
Interoperability is a major challenge in current healthcare systems. It brings big hope for data exchange, but also raises some concern about patient safety. We study the wireless updating of modern infusion pumps and demonstrate the possible flaws in this process. Through analyzing data on drug limit libraries (DLL) versions in one hospital we could identify the delays in distributing DLL updates and the impact these delays might have on patient safety. We found that 31% of all started infusions had used outdated DLL versions, and 22.6% of all alerts were triggered by outdated DLLs. These findings suggest that clinical and operational stakeholders in healthcare systems must address the unreliable interoperability of medical technologies such as seen on infusion pumps. The impact of information inconsistency across healthcare systems might result in use error which would impair patient safety.

Introduction
Medical devices’ advancement in automation and interoperability is playing a significant role in the way they are integrated into the drug delivery system. In order to make all component synchronized in the integrated system, which includes the human operator and the medical devices, it is important to have very reliable communications between all of them. Devices that are not synchronized may cause inconsistency (or, from the operator’s viewpoint, operational failure) which in turn might translate into use errors and distrust of the technology. Such distrust likely would lead to low adoption rate of using the technology and also undesired workarounds1, 2.

Infusion pumps are among the technologies that have advanced interoperability in the clinical setting. Modern infusion pumps models have built-in dose error reduction systems (DERS) and drug limit libraries (DLL) that can alert users when a drug is programmed outside of the preset limits, and they have shown to be effective in reducing potential infusion errors3. These functions were designed as a safety feature that should ease and standardize the way clinicians program the infusion pump. Moreover, hospitals often make changes to drug limits to reflect evolving clinical needs, and, thus, would publish new versions of DLL from time to time. Such updates on DLL are done wirelessly nowadays without the need to perform updates of the DLL file on each individual pump manually with a cable. Ideally, in order to standardize clinical workflow and avoid confusion, all infusion pumps in the hospital should have the same version of the DLL especially with such wireless automation.

Objective
The objective of this study is to raise clinicians’ awareness of the issue of inefficient infusion pump DLL updates, and what the impact may be. The problem of delayed updates on pumps may have been known by some but not so much about its magnitude, its impact on drug limit compliance and the potential risk of patient harm4, 5. Following the authors’ prior descriptive work on the issue of long update delays of DLL in several hospital systems using merely infusion alert data5, we took a step further to collect pump status and all infusion detail reports in addition to the alert data from one of our collaborating hospitals. By cross checking the numbers in each of these different types of report, we would also be able to discover in-depth the spread and impact of such issue on infusion drug alerts, potential patient harm and its contribution to nursing workarounds.

Materials and Methods
Data
The data for this analysis came from three infusion pump data sources. One is drug alert data from the Infusion Pump Informatics (IPI) System maintained and supported by the Regenstrief Center for Healthcare Engineering (RCHE) at Purdue University6. Another is the pump status report which can be generated by a vendor pump management software. The other data source is the detailed all infusion report, available through the vendor analytics software tool,
which contains each infusion start of the dates of interest, with or without triggering drug alerts or alarms. Each report
type is explained as follows.

Infusion Pump Alert Data. All IPI community member hospitals share pump alert data by periodically uploading them
to the IPI System on CatalyzeCare, a Hubzero©-based cyber infrastructure. An infusion alert happens when a
programmed value violates the preset limits for the drug and its clinical use, and the clinician can decide to accept the
alert by cancelling or reprogramming the infusion or override the alert. As of late 2015, the system contained more
than 21 million infusion pump drug alerts from more than 120 hospitals across the U.S. Even though the IPI data
contains alert data from multiple pump manufacturers, we focused our analyses using data from one of them. Each
line of alert includes detailed information such as drug name, profile name, drug limit library version used, infusion
type, programmed value, drug limit violated, time stamp of infusion attempt, etc.

Pump Status Reports. Pump status reports are available through the vendor provided pump management software tool
which is different from the analytics suite. A pump status report is a snapshot of all the pumps in the network. For
each pump it details the location (facility), model, pump serial number, activated DLL (i.e., the version all pumps
should be using at the moment), DLL version on the very pump, any pending DLL version on the pump (i.e.,
downloaded DLL file but yet to be installed), last communication time stamp (between the pump and the server) and
profile last observed.

All Infusion Detail Reports. As the name suggested, all infusion relevant events are included in an all infusion detail
report, such as infusion started, restarted, stopped, paused, alerted, resolved, time stamp, whether or not the event is
under guardrails system, type of infusion (continuous, bolus, etc.), actual infusion duration seconds, programmed
duration seconds, volume to be infused, dosage, and infusion rate, etc. An infusion delivery is actually composed of a
sequence of infusion events which share the same infusion ID. As show in the report, usually an infusion
administration starts with an “infusion started” event and ends with an “infusion stopped”. In summary, the all infusion
report lists information regarding when an infusion event happened, for what purposes it happened, how the infusion
was programmed (using basic mode or with guardrails), and the detailed infusion settings. This allows us to study not
only the infusion alerts generated by various DLL versions but also the bigger picture of all infusion starts that did not
generated any alert.

Methods

We used the alert data of an inner-city hospital of 315 beds in the Midwest from July 1st to September 18th, 2015 and
the DLL version information indicated in each alert to show how many and which DLL versions exist in the infusion
drug alert. There were 680+ pumps (i.e., the main units with memory) in the facility. We also collected the hospital’s
pump status reports and all infusion detail reports for the same period of time. Since the pump status report is a
snapshot of each pump’s status in the fleet, we requested that the pharmacist in charge to send one to us daily if at all
possible in order for us to keep track of the status change of each pump and the patterns of the changes. Our partner
hospital had three DLL updates during that period of time. Note that an update cycle is defined as the time from the
first day of a new published version of DLL to the day before the publication of the next DLL version. We then
classified, on that specific reporting day, each pump as “current” or “outdated” by comparing the DLL version on
each pump with the activated DLL version. We further examined the severity of the DLL update problem in all
infusions recorded in the all infusion detail reports.

Results and Discussion

DLL Update Rate by Alerts

Figure 1 illustrates the phenomenon that alerts could be generated by outdated DLL versions. Each color represents
the alerts trigger by a DLL version, and each column is a day. In this hospital, alerts were seen to be generated by
seven different DLL versions within an 80-day window. As the figure shows, once a DLL update is pushed out to the
pumps, alerts by the new version started to happen (as a new color starts to show in the bars). However, the more
troubling is the fact that some alerts were still being triggered by updated DLL versions as their colors bleed into more
recent days. Should the drug limit settings be different in the new versus the old DLL versions, clinicians may be
getting false alerts (i.e., those that should not have happened) or missing alerts (i.e., those that should have happened).
Figure 1. Percentage of alerts by DLL versions.

DLL Update Rate by Pump Status Reports

We tracked each individual pumps and their status using the pump status reports. Figure 2 shows the distribution of the DLL versions on each day of the reporting. Each time a new color appears in the vertical bars it indicates that a new DLL version has been released and has been installed on some of the pumps. However, as shown in the figure, more and more colors show up in the bars as time goes. It means that the number of DLL versions kept increasing every time a new one is released, i.e., the old versions never got replaced on some pumps during this period of time.
Figure 2. The number of pumps (shown) of each DLL version. The length of each color on the bar indicates the percentage of such version found on that day.

We further showed the effect of the pumps using outdated DLL versions by combining the alert counts and the status of each of the pumps on these days. Figure 3 shows the correlation of the m: the diamond markers are percentages of pumps with outdated DLL based on pump status reports, and the square markers are the percentages of alerts generated by outdated DLLs. Naturally, the more pumps with outdated DLL versions the more alerts would be generated by outdated DLLs especially shortly after a new version of DLL was published and pushed out to all pumps in the network. If the probability of alert generated per pump and the utilization of pumps be equal on all, these two percentages should remain very close to each other on any given day. One possible explanation of the gaps between the diamond and square markers on some days may be that some pumps were not being used as often as other were. In other words, it is reasonable to speculate that some pumps were used and/or turned over much more often thus they were provided more opportunities to get updated and be generating alerts. More analysis is needed for us to understand this phenomenon better.
Impact of Using an Outdate DLL Version on the Pump

We extracted all events of infusion started from the all infusion detail reports during this period of time and investigated how many of started infusions were on pumps with outdated DLL versions. The results were further categorized by infusion type (Table 1). It shows that on average almost one third of all infusions started with an outdated DLL version, and, among all infusion types, bolus infusions had the largest incident rate (44%). The results here also indicate that over 30% of all attempts of infusion starts were at risk of not equipped with the most up-to-date version of the DLL on the pump. Recall the DLL analysis on infusion drug alerts seen earlier in this section, the results of infusion started events showed us that drug alerts do only represent a small portion of all that might have been “risky” infusions.

Table 1. Numbers and percentages of infusion started events performed on pumps with outdated DLL versions from July 1 to September 18, 2015.

<table>
<thead>
<tr>
<th>Infusion Type</th>
<th>Total # Infusion Started Events</th>
<th># of Infusion Started Events with Outdated DLL Versions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>28150</td>
<td>10011 (36%)</td>
</tr>
<tr>
<td>Bolus</td>
<td>10432</td>
<td>4548 (44%)</td>
</tr>
<tr>
<td>Fluids</td>
<td>31640</td>
<td>8288 (26%)</td>
</tr>
<tr>
<td>Intermittent</td>
<td>32507</td>
<td>9147 (28%)</td>
</tr>
<tr>
<td>Total Number</td>
<td>102729</td>
<td>31994 (31%)</td>
</tr>
</tbody>
</table>

Conclusion

In this study we exhibited and verified long delays of infusion pump drug limit library updates using multiple data sources from one hospital including pump drug alerts, pump status reports, and all infusion detail reports, all available through vendor provided software tools. In this specific case study of one hospital we saw that on average more than 50% of the pumps were using outdated DLL on reporting days, 31% of all started infusions had used outdated DLL versions and 22.6% of all alerts were triggered by outdated DLLs. This has great implication of potential inconsistency
of pump response to the operator, and the possibility of patient harm due to incorrect drug limit settings (such as when an error made in the DLL was not able to be corrected in time for patient care). Outdated DLL triggered false alerts may cause confusion of the nursing staff and attribute to alert/alarm fatigue problem well known in the healthcare setting while missed alerts may allow inappropriate infusions to be administered. It is also very likely that a nurse would choose to use “basic mode” of infusion, i.e., electing not to apply drug limit settings to the infusion, when s/he cannot find the drug due to not having the most up-to-date DLL on the pump. Such workaround or deviation from the best practice of using the limit settings (i.e., noncompliance), is highly undesirable because of its potential to cause patient harm should a programming error occurs.

Our analyses provided different ways that a hospital can utilize pump informatics to track the effectiveness of pump updates and monitor their progression and impact. We believe it is crucial in ensuring consistency in nursing workflow, thus reducing unnecessarily workarounds, and most importantly, patient safety in hospital care. Our future work will include more in-depth analyses on missed or false alerts due to outdated DLLs.

References

Evaluation of a Digital Companion for Older Adults with Mild Cognitive Impairment

George Demiris, PhD, FACMI¹, Hilaire J. Thompson, RN, PhD, FAAN¹, Amanda Lazar, PhD², Shih-Yin Lin¹
¹University of Washington, Seattle, WA, USA; ²Northwestern University, Evanston, IL, USA

Abstract

Study Objective: The purpose of this study was to examine the feasibility of a digital companion system used by older adults with mild cognitive impairment (MCI). We utilized a commercially available system that is comprehensive in its functionalities (including conversation ability, use of pictures and other media, and reminders) to explore the system’s impact on older adults’ social interactions, anxiety, depressive symptoms, and acceptance of the system.

Study Design: We conducted a three-month mixed methods evaluation study of the digital companion.

Results: Ten female community-dwelling older adults (average age 78.3 years) participated in the study. Overall, participants utilized the tool regularly and appreciated its presence and their interactions. Participants scored higher at the end of the study in cognition and social support scales, and lower in presence of depressive symptoms.

Conclusion: Findings indicate the feasibility of a digital companion for people with MCI and inform the need for additional research.

Introduction

In recent years, various information and communication technology (ICT) applications have been developed in order to engage older adults with cognitive impairments in entertainment or reminiscence therapy or to simply provide a forum for interactive and tailored engagement. The types of applications available can be grouped into those that mitigate specific impairments of their target users, such as motor impairments or memory deficits, and those that take advantage of and aim to maximize continuing abilities. Technology has been used accommodate motor impairments and aid interaction through commercially available gaming devices, touch screens, and prototype devices. The MINWii videogame employed a Wii remote as an input device to accommodate motor impairments of users who might not otherwise be able to play the game.¹ ¹ Shik, Yue, and Tang (2009) used headphones and amplifiers to accommodate hearing impairments and projectors to magnify photographs to accommodate vision impairments.³ Other systems may gather, for example, materials from the users’ daily activities using technology, such as GPS, cameras, and audio recorders, to compensate for memory deficits).⁴ Tools have been designed to alleviate caregiver strain originating from the need to repeatedly provide details to help someone with dementia recall facts about events during casual reminiscence.⁵ Yamagami and colleagues used a video at the beginning of a session with the ICT to remind participants how to use tools. The participants then showed staff how to use these tools, reversing the roles of staff as helper to resident as helper. By using a video instead of having staff instruct the residents, this role reversal was possible.⁷

Certain skills, such as sensory awareness, musical responsiveness, and emotional memory, have been labeled “continuing abilities” in dementia, as they are thought to persist after others have been compromised by the disease.⁸ Sensory awareness is the response to various forms of stimuli (e.g., visual, audio, tactile). Musical responsiveness refers to the strong responses people with dementia can have to music. Emotional memory refers to the ability to experience rich emotions. Technology is commonly used to encourage and draw out these continuing abilities. Such
as through the display of movies, photographs, and audio that elicit sensory or emotional responses. Many systems that promote reminiscence with people with dementia use technology solely for this purpose. Media are used as triggers to prompt a positive response in the form of interactions or improved mood. Reminiscence systems that focus on emotional memory appeal to something personally relevant to the individual user. This ranges from materials that are personal only in that they are from the general era in which one grew up to artifacts related to individual interests or even objects that belonged to the individual.

ICT tools that integrate one or more of the functionalities described above can provide an integrated system that is meant to not only engage older adults in specific activities and tasks but also create an opportunity to address social isolation and loneliness. The effects of loneliness and social isolation are well documented for older adults; a recent meta-analysis of 148 longitudinal studies reported a 50% reduction in the likelihood of mortality over a period of 7.5 years for individuals with strong social relationships. Social isolation is negatively associated with health status and health-related quality of life of older people with effects magnified among older adults with dementia. ICT applications that address loneliness and social isolation often have anthropomorphic or animal features in order to generate a likeness to a friend or pet, and are referred to as digital companions. An older adult can interact with such a digital companion both at pre-programmed times (for example, when reminders or other messages are sent) as well as at the discretion of the user. Conversations can range from simple exchange of messages to sophisticated discussions depending on the system’s features and the level of machine learning algorithms, artificial intelligence or human response involved at the other end.

Digital companion tools are also referred to as Embodied Conversational Agents (ECAs). ECAs interact with users through verbal and non-verbal behavior cues such as prosody (pacing and intonation) and hand movements. A small number of studies have emerged examining the use of digital companions or ECAs. Bickmore et al. (2013) created a virtual laboratory to study users’ reactions. Elderly participants interacted with an ECA acting as an exercise coach. Results showed that users who interacted with an ECA that used variable dialogue exercised significantly more than those interacting with an ECA with non-variable dialogue. Vardoulakis et al. also investigated the use of an agent to provide social support and wellness counseling for older adults. A system was constructed that allowed research assistants to control an ECA placed in an older adult’s home in real time. Qualitative analysis of the interactions identified multiple topics that older adults liked discussing with the agent and general design principles towards building future companion agents for older adults.

While these studies have introduced insight into the potential of digital companion tools, the vast majority of them have been tested within laboratory settings, where older adults were unable to interact with the system over a long period in a naturalistic setting, or only tested with one user only. Furthermore, these studies are not longitudinal and for the most part rely on assessment of the system when it is used once or for a limited number of times. Finally, most of these digital companions either provided simply opportunity for conversation, or focused on one aspect of interaction by showing pictures or other media or served a single purpose (for example, exercise coaching or reminding of upcoming events and medication).

The purpose of this study is to examine the feasibility of digital companion systems in real world settings used by older adults with mild cognitive impairment in their natural environment and for a longer period of time. For this purpose we utilized a system that is comprehensive in its functionalities (including conversation ability, use of pictures and other media, and reminders) to explore the system’s impact on older adults’ social interactions, anxiety and depressive symptoms, and participants’ acceptance of the system.

Methods

We conducted a three-month mixed methods evaluation study of an existing commercially available digital companionship device. We recruited community-dwelling older adults.

Recruitment

Older adults were recruited through 1) posted flyers (at sites such as hospitals, memory clinics, day health programs, memory care groups and retirement communities with permission of those sites); 2) flyers sent to people in memory groups with the permission of memory group leaders; 3) word of mouth and snowball sampling; and 4) with the
permission of memory group and day health program leaders, we approach participants directly and told them about the study using language similar to that on the flyers. Older adults interested in participating contacted a member of the research team. We utilized a phone screening form that also included the administration of the Memory Impairment Screen-Telephone (MIS-T) instrument, a phone based instrument to screen for cognitive impairment in older adults. The maximum score is 8. If an older adult scored 6 or below, this indicated the possible presence of cognitive impairment, and then a member of the research team met with them to assess them further for eligibility. Additional eligibility criteria for older adults included: 1) be able to see and hear well enough to interact with the device to some degree (as assessed by researcher), and 2) reside in the Seattle metropolitan area. Participants were excluded if they were 1) unwilling to be audio recorded 2) unable to speak English. Family members were identified by older adults who agreed to be in the study. If an older adult chose to identify a family member to participate in the study, a member of the research team contacted them to provide information about the study and recruit them if they were eligible and interested. Interested and eligible persons were provided with information about study procedures, risks and benefits and asked to explain their understanding back to the research staff. All study participants provided written informed consent.

Intervention
The virtual pet companion is displayed on a tablet as part of the GeriJoy service provided by a company called “care.coach.” It interacts with the older adult through voice and expression. It stays plugged in to a dock by the older adult’s bedside or other preferred location in their residence. On the other end of the device is a trained staff member of a 24/7 call center who listens to the older adult and types responses that are then converted to audible speech. Thus, the “pet” is fully controlled by a human who uses both scripted text and unscripted spontaneous exchange for all the interactive sessions with the end user. The call center staff summarize their interactions with the older adult and send a log to a family caregiver. The caregiver can also send pictures to the device to share with the older adult (see Figure 1). In this study, we provided older adults with the option to share this log with a family member, but it was not automatically transmitted to anybody besides the research team. The device allows the human on the other side of the tablet to see the older adult in their room (when the pet is “awake,” signified by its eyes being open) to determine if the older adult is interested in engaging in conversation and to add a dimension to the interaction. In this study, we allowed the older adult to opt-in to having this video feature. This application supports all functionalities of ICT tools for dementia and MCI described earlier, namely addressing cognitive deficits by providing reminders and prompts (based on a pre-defined script that is used by the call center), use of pictures and multimedia to facilitate sensory awareness and memory support and it furthermore facilitates interactive communication ranging from small talk to any topic of concern to the older adult.

Study Procedures
Older adults were given a digital companionship device to use for three months. Semi-structured interviews were conducted at baseline, midpoint, and three months. The following instruments were administered at baseline and three months:

-Montreal Cognitive Assessment (MoCA): A 10 minute test covering 8 domains to screen for cognitive impairment in older adults. Higher scores indicate better cognitive function. Scores can be used to rate cognitive impairment as mild, moderate, or severe. The sensitivity and specificity of the MoCA for detecting early dementia are 100% and 87% respectively.17
- **MOS Social Support Survey (MOS-SS):** A 19-question test to measure social support. The instrument has been tested extensively for reliability (with Cronbach’s Alpha greater than 0.91) and construct validity. There are four subscales to the MOS-SS: Emotional/Informational Support, Tangible Support, Affectionate Support and Positive Social Interaction. Each of the subscales are scored 0-100 with a higher score indicating more support.

- **Patient Health Questionnaire PHQ-9:** A 9-question test to measure depressive symptom severity. This tool is extensively tested for reliability and validity including in use with older adults with MCI and early dementia. A higher score indicates more depressive symptoms.

- **Generalized Anxiety Disorder GAD-7:** A 7-question test to measure overall anxiety (general anxiety disorder. This instrument has been found to have high levels of reliability, criterion, construct factorial and procedural validity (Spitzer, Kroenke, Williams, & Loewe, 2006). A higher score indicates higher levels of anxiety.

The following instrument was administered at 3 months only:

- **Comfort from Digital Companion Animals Scale:** An 11-item questionnaire to assess participants’ attachment to the digital companion, which was modified from Zasloff’s Companion Animals Scale.

### Results

A total of ten older adults participated in the study. All participants were female and their average age was 78.3 years. Eight of the ten participants completed all study procedures. Two withdrew before the midpoint interview. One of the participants who withdrew from the study had more advanced cognitive impairment and felt it was too difficult for her to trouble shoot on her own when the device was not working (e.g. not being able to connect to the internet, the device not charging properly, etc); the other participant who withdrew from the study had a hard time getting the device to work for a few days (mainly due to issues with her new Wi-Fi) and felt that she needed a deeper level of interpersonal connection than what the digital companion was able to offer. Only one of the participants chose to invite a family member to participate in the study.

Table 1 summarizes demographics and baseline use of technology. Table 2 compares survey responses at baseline and at study exit. Although there were not enough participants in this pilot study to do robust statistical analyses, we did assess change from baseline in each of the instruments, noting reduced depressive symptoms. There was an increase noted in both the MoCA and the MOS-SS, with the largest benefit seen, not surprisingly, in the positive social interaction subscale. There was a small increase in anxiety at study exit. We explored the association between change in anxiety and attachment to the digital companion (recognizing that such a sub-analysis is weakened by our small sample size). Interestingly, those with higher attachment to the pet had baseline GAD of 3.75, increasing to 5 at end while those with lower attachment (not very/not attached) actually had a decrease on the GAD from 1.75 to 0.75. Table 3 includes the level of attachment that participants had at study exit.
Table 1. Demographics and Baseline Use of Technology.
All data are reported as % (n) unless noted.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pre-Test (Baseline n=10)</th>
<th>Post-Test (n=8)</th>
<th>Average individual change T1 to T2 (n=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean; range)</td>
<td>78.3 Years (68-89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female Gender</td>
<td>100% (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>90% (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native American</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever owned a pet</td>
<td>100% (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comfort Using Technology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Uncomfortable</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somewhat Uncomfortable</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somewhat Comfortable</td>
<td>70% (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Comfortable</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of Technology for Leisure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly Dislike</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dislike</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>20% (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Like</td>
<td>50% (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly Like</td>
<td>10% (1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison of Survey Responses at Baseline and at Study Exit after Using GeriJoy System for 3 months. Data are reported as mean (SD) unless noted.

<table>
<thead>
<tr>
<th>Measure (Tool)</th>
<th>Pre-Test (Baseline n=10)</th>
<th>Post-Test (n=8)</th>
<th>Average individual change T1 to T2 (n=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognition (MOCA)</td>
<td>21.9 (7.3)</td>
<td>23.5 (3.3)</td>
<td>+.13</td>
</tr>
<tr>
<td>Social Support (MOS SSS)</td>
<td>69.9 (14.5)</td>
<td>72.6 (15.9)</td>
<td>+1.36</td>
</tr>
<tr>
<td>Subscale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional/Informational</td>
<td>65.6 (23.1)</td>
<td>69.5 (22.0)</td>
<td>+1.17</td>
</tr>
<tr>
<td>Tangible</td>
<td>72.5 (17.2)</td>
<td>71.1 (25.9)</td>
<td></td>
</tr>
<tr>
<td>Affectionate</td>
<td>65.0 (30.9)</td>
<td>67.7 (30.7)</td>
<td>+2.08</td>
</tr>
<tr>
<td>Positive social interaction</td>
<td>66.7 (21.9)</td>
<td>77.1 (20.1)</td>
<td>+6.25</td>
</tr>
<tr>
<td>Anxiety (GAD-7)</td>
<td>2.5 (1.7)</td>
<td>2.9 (2.6)</td>
<td>+.13</td>
</tr>
<tr>
<td>Depressive symptoms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9</td>
<td>3.5 (2.1)</td>
<td>2.5 (1.8)</td>
<td>-.88</td>
</tr>
<tr>
<td>Difficulty to do things at work, home, get along with other people if problem noted in general PHQ-9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not difficult at all or N/A</td>
<td>80%</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>Somewhat difficult</td>
<td>20%</td>
<td>25%</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Participant Responses at Study Exit pertaining to the digital pet. Data are reported as mean (SD) unless noted.

| Table 3. Participant Responses at Study Exit pertaining to the digital pet. Data are reported as mean (SD) unless noted. |
|---|---|
| Comfort from Digital Companion Animals Scale | Post-Test (n=8) |
| | 27.8 (7.9) |
| How attached were you to this digital pet? | |
| Very attached | 25% |
| Attached | 25% |
| Not very attached | 37.5% |
| Not attached | 12.5% |

At baseline, participants expressed interest and curiosity as to how their interactions with the digital companion would evolve over time. As one participant stated “it’s just an adventure; like, I like going to the secret film festival because we don’t know ahead of time what movies we’re going to see.” Another participant stated “I will have something to look forward to play with to do, something with each day.” When asked why they agreed to participate, few participants indicated they were curious and wanted to try new things out, others talked about helping out with research projects that can help future generations and one participant specifically mentioned the inability to have and maintain a real pet as the reason: “I am very sorry that I don’t have a pet here because I’ve always had pets. They’re an integral part of the way I relate to life.”

In the exit interviews, participants overall discussed being satisfied with the digital companion and having enjoyed their interactions. When asked to identify system strengths, participants talked about having a remote friend who is always available and ready to talk when one is lonely, the significance of the reminders that the digital companion would provide (for medication adherence or dietary restrictions) and the ease of mind that a remote “friend” provides by “checking on you regularly and asking if things are ok when [the pet] hears a strange noise or sees something strange.” One participant also appreciated the visual communication whenever the digital companion would show pictures or little hearts: “I would have it right next to my chair, the hearts would go. When people would come, they’d say, “What’s that?” I said, “Those are just little love notes. [chuckles].” Another participant also showed appreciation for times when the pet would use pictures to entertain the participant: “He found a birdbath with some crows in it on Google. He was always very good-very sensitive to what I was trying to do.”

One weakness identified by participants was the fact that the interaction was strictly verbal or visual. As one participant pointed out “I wish it was something you can hug, that you can touch, like a real cat or dog.” While numerous participants commented that the communication was at few times problematic with the pet either having a limited vocabulary, being repetitive in its questions, not remembering details of previous conversations or asking questions at inappropriate times (e.g., when the participant was taking a nap or watching TV), other participants commented on the sensitivity that the “pet” showed as a conversation partner in engaging only in conversation topics that participants wanted to discuss at a given time.

**Conclusion**

Our study findings indicate the feasibility of a digital companion tool for older adults with MCI. Overall, participants utilized the tool regularly and reported appreciating its presence and the time they interacted with it. Participants scored higher at the end of the study in their assessment of cognition and social support and lower in presence of depressive symptoms. Anxiety overall was higher at the end of the study, but upon closer investigation, anxiety increased for those who felt attachment to the digital pet (potentially because they had to give up their digital companion).

This is an exploratory feasibility study with a small sample size and no control or comparison group, which clearly limits the generalizability of our findings and does not allow for identification of confounding factors or natural trends. Additionally, participants were selected from only one geographic region. The sample had limited diversity.
in terms of race and ethnicity and no diversity in terms of gender, as all participants were female. Additionally, only one participant chose a family member to participate with them in the study. This may be due to privacy concerns or a perception that participating in the study may become burdensome to family members. The feasibility of family involvement needs to be further investigated. The system was deployed for three months for each participant, and while this time frame is the longest documented for these types of systems, it may still be too short to fully examine adherence to the intervention. In spite of these study limitations, our findings highlight the potential of digital tools to provide companionship, reminders, and potentially health behavior coaching for older adults with MCI. Future studies need to be based on experimental design and include a control group in order to understand potential confounding factors and natural trends, and should furthermore address how the use of such a tool affects clinical outcomes as well as ways to meaningfully engage family members or other stakeholders. Design considerations need to be assessed to determine the extent to which such a tool meets older adults’ preferences and needs. Finally, ethical considerations around attachment to a digital tool and unintended consequences in cases where the tool is no longer functional or available need to be explored.

Acknowledgment

This study was supported in part by the NIH National Institute for Nursing Research Aging and Informatics Training Program Grant Nr. T32NR014833 at the University of Washington.

References


Using Best Practices to Extract, Organize, and Reuse Embedded Decision Support Content Knowledge Rules from Mature Clinical Systems

Spencer J. DesAutels, MLIS¹, Zachary E. Fox, MSIS¹, Dario A. Giuse, Dr.Ing², Annette M. Williams, MLS¹, Qing-hua Kou, MS¹, Asli Weitkamp, PhD², Neal R. Patel, MD²,³, Nunzia Bettinsoli Giuse, MD¹,²

¹ Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN; ² Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; ³ Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN

Abstract

Clinical decision support (CDS) knowledge, embedded over time in mature medical systems, presents an interesting and complex opportunity for information organization, maintenance, and reuse. To have a holistic view of all decision support requires an in-depth understanding of each clinical system as well as expert knowledge of the latest evidence. This approach to clinical decision support presents an opportunity to unify and externalize the knowledge within rules-based decision support. Driven by an institutional need to prioritize decision support content for migration to new clinical systems, the Center for Knowledge Management and Health Information Technology teams applied their unique expertise to extract content from individual systems, organize it through a single extensible schema, and present it for discovery and reuse through a newly created Clinical Support Knowledge Acquisition and Archival Tool (CS-KAAT). CS-KAAT can build and maintain the underlying knowledge infrastructure needed by clinical systems.

Introduction

Vanderbilt University Medical Center’s (VUMC) long history of innovation in biomedical informatics has led to the creation of unique and increasingly complex clinical decision support (CDS) over multiple decades. These CDS resources are embedded in the many information systems within the institution at the point of care. This paper presents VUMC’s efforts to externalize this varied content in a systematized manner, leading to the creation of a comprehensive CDS schema and knowledge acquisition and archival tool.

At VUMC, CDS is tightly integrated with the system in which it is implemented. Horizon Expert Orders (HEO), the computerized physician order entry (CPOE) system originally developed at Vanderbilt as WizOrder,¹ presents CDS ranging from allergy checking and drug-drug interactions to discharge instructions and complex protocols. Within HEO, CDS is written in a locally-developed scripting language called Vanderbilt Generalizable Rules (VGR) that allows user screens to be overridden with simple display templates and HTML files.²⁻⁴ Within StarPanel,⁵ Vanderbilt’s electronic health record (EHR) system, built-in CDS visually presents relevant information derived from the patient’s data. These display resources are called Indicators⁶ because they rapidly indicate critical information about the patient. Indicators are created by professional developers in the high-level programming language Perl. Other systems such as the outpatient order entry system, VOOM, similarly integrate unique CDS directly into the system with which providers interact. This approach has led to the implementation of CDS through diverse languages and mechanisms.

Support for Horizon Clinicals, the suite of clinical systems¹ used by VUMC which includes HEO, HED (the inpatient nurse documentation system), AdminRx (the barcoded medications administration system), and HMM (the pharmacy management system) will be discontinued. In light of this, VUMC leadership created the Clinical Systems 2.0 initiative⁷ to replace the clinical systems with a new suite of products. The part of this institution-wide effort presented in this paper is focused on transitioning select CDS content knowledge from the current systems and implementations to representations that retain clinical significance and intent and can be implemented in the new clinical systems. The Center for Knowledge Management (CKM) team was approached directly by the Chief Informatics Officer and asked to apply its unique set of skills towards this project in March 2015.

The CKM team at VUMC aids in supporting the information-strata of the organization through innovative processes of knowledge acquisition and organization for the purposes of archiving and reuse. As such, the CKM team is well

504
situated to work on this project given our close linkage to biomedical informatics and our expertise with metadata schema development, long standing provision of actionable evidence in the clinical setting, relational database and interface creation, content curation, and extraction and identification of linkages between disparate information sources.9,10

Preliminary meetings with leadership, CDS content creators, and stakeholders further demonstrated both the scope of the institution’s need and that CKM’s skills could easily, and immediately, be directed towards managing the existing content. For example, as evidence has been a critical part of creating and updating CDS resources, the CKM team recognized the opportunity to extend and update this evidence alongside the executable content. Further, production databases hold both content that is currently used and content that had been intentionally deactivated, presenting an opportunity for archiving and future retrieval. Additionally, the meetings further aided CKM in fostering collaboration with HealthIT’s Knowledge Engineering team at VUMC. The knowledge engineers, using their relationship with clinicians and end-users of CDS content, helped shape the approach the CKM team used to refine tool development, and eventually, content validation. The meeting outcomes, along with CKM’s long history of expert evidence provision in complex environments,11 made ongoing content maintenance another key goal for the project.

Implementation Background and Methods

Synthesizing the institutional need, the preliminary data we gathered from meeting with experts and stakeholders, and our own skills, we arrived at the following goals for a unified approach to managing knowledge around clinical decision support: extract content and medical context from application-specific representations, preserve complexity, identify and update evidence to support current practice, enable cross-system search for the discovery of previously obscured linkages, validate content, and aid in the process of making decisions on active versus archival retention. These goals would be tested over the entirety of active content from two types of CDS: VGRs and Indicators. The rationale behind this implementation was dictated by our goals and the relevant literature.

Our first and most critical goal was to extract content directly from the multiple applications in which it was created and represent it in a unified way. As this initiative was started well in advance of new clinical systems being implemented, VUMC did not have a replacement suite of systems selected, and it became a major design goal to keep the extracted knowledge system-agnostic. We therefore focused first on standardization efforts such as Arden Syntax, Knowledge Artifacts, and GELLO that show the value of representing CDS in a generic way using a consistent schema. In Arden Syntax, basic descriptions of content are standardized, while data to support the logic are kept entirely application-specific, which means the decision support must be mapped to each application to be reused.12 Knowledge Artifacts solve this challenge through an XML representation of logic using the Virtual Medical Record (VMR) domain analysis model of the HL7’s Reference Information Model (RIM) to promote interchange and interoperability.13,14 GELLO likewise uses the RIM to enumerate all possible data types and type extensions.15 The CKM team conducted a thorough literature search of CDS standards, schemas, taxonomies, ontologies, and development methods and considered the aforementioned standards as ways of creating implementation-independent representations of the CDS resources at scale, while allowing for extensibility to meet evolving needs.

Another goal of the CKM team’s effort was to facilitate unambiguous human understanding of complex rules while ensuring ease of use. Local CDS that implements a protocol or guideline is complicated, and to allow users to truly understand that protocol we had to correctly capture all of its intricacies. The Clinical Decision Support Consortium created a model for representing guidelines that uses multiple layers of increasing complexity from text to semi-structured, structured, and executable,16 similar in nature to the multiple representation formats of DeGeL.17 This allows consumers of the representation to pick the level that best suits their needs, potentially increasing use.16 CKM similarly chose to create a tiered approach that allows both complexity and simplicity depending on user preference.

A chief concern of the CKM team was enabling discoverability of resources in as many ways as possible. As we were starting from a collection of software artifacts, some understanding of methods from software reuse were beneficial to our efforts. Software reuse is a software engineering concept that starts from a piece of code that provides an important function and creates a searchable abstraction. The goal of this process is that useful snippets of code can be reused as much as possible, saving future development time. While not quite the same as our goal of recreating rather than reusing these resources, software reuse provides four key components of reusability: abstraction, concise description, specificity enough to distinguish between descriptions, and a clear understanding of...
relationships to other artifacts. Further, software reuse literature describes eight classical methods of abstraction that can fulfill these needs, whether for reuse or adaption.

Other work in biomedical informatics has further described the content and context of decision support beyond standards. Decision support taxonomies in the literature describe the functions decision support can provide, based on empirical observation. Greene presented a general schema based on purpose and design. This body of work provided context for thinking about discoverability specific to CDS.

It was important to the team to automatically extract as much as possible from the existing disparate systems and present the resulting information in a single repository for users to find and view CDS. Sittig et al. defined four tools needed to effectively manage clinical knowledge: an external repository of clinical content, a collaborative space to develop content, a terminology control tool, and a mechanism for end users to provide feedback. Our focus was on creating this external repository through iterative feature development based on stakeholder requirements.

Finally, it was critical to identify clinical owners, validate our work, and prioritize the CDS content to move into the new systems. Validation and decision-making on CDS prioritization was provided by the Knowledge Engineering team in partnership with clinical content owners. Further, the first step in validation was ensuring the schema was robust through testing across a varied set of highly used and logically separate resources as a representative sample. Reference documentation was used to aid in application of the schema to each clinical system. Knowledge Engineering team members were also encouraged to contribute to populating any content fields that their expertise could provide.

This intervention took place within the clinical systems of VUMC’s tertiary-care teaching hospitals, which serve middle Tennessee and the surrounding areas, as well as inpatient and outpatient facilities, all of which use StarPanel as the EHR.

Results
In all, the CKM team extracted 1024 Indicators (965 active; 59 inactive), 3147 VGRs (1671;1476), 1875 display templates (945;930), and 1951 HTML pages (705;1246) for representation in our in-house created web-based tool. This tool (Figure 1), named the Clinical Support Knowledge Acquisition and Archival Tool (CS-KAAT), allows all the above knowledge to be viewed and edited. CS-KAAT uses a MySQL database with a PHP front end. Access is restricted by login, and user role enables various functions from access to editing specific fields. Codified information is collected automatically via scripting and validated by human agents. If the automatically extracted data needs to be modified, the new values for that field are clearly marked as edited, allowing end users to differentiate which party to trust for the accuracy of the information. Changes to production databases are captured periodically and automated content extraction and manual reconciliation are performed on all new or modified resources to ensure that content in CS-KAAT remains current with production systems.

Figure 1. Clinical Support Knowledge Acquisition and Archival Tool (CS-KAAT).
Our schema provides 70 elements informed by multiple literature sources and modified to fully describe CDS content at VUMC. Content is described along central facets with controlled vocabularies that enable comparison. This method of software abstraction allows high precision and recall, while providing transparent and easily understood access to users.\textsuperscript{19} The schema driving CS-KAAT can be divided into important components – for instance, divisions taken from the high-level concepts of the GEM model\textsuperscript{27} – to present the elements (Figure 2). Identity elements uniquely identify a resource. Developer elements are the names and roles of people involved in the creation of a resource, and Testing elements identify editorial oversight and modifications. Purpose elements define the reasons a resource was implemented. Intended Audience elements identify the clinicians impacted by a resource, while Target Population elements delineate patients affected by the resource. Knowledge Components elements store the knowledge from applicable evidence, local practice, and implemented logic and display. Finally, Revision Plan elements ensure currency of the resources.

The schema applies to decision support of any type by mapping the features of that particular type of content to the appropriate schema elements. Mapping to each system allowed us to see that the schema element ‘Category of Intervention’ would receive the same value for all Indicators, but individual VGRs could have values for this element different from each other. From this realization, schema elements are mapped to the most appropriate of four levels to fully describe the CDS resource: system, group, resource, or resource part.
System-level elements apply equally to each resource from the same system. For example, the ‘Locus of Control’ is “system” and ‘Rights’ are “Copyright VUMC” for every VGR, and are therefore applied once to the entire system.

Within CS-KAAT, all users are able to create a generic group and can include any resource as members of their newly created group. This allows users to collocate resources for specific interests or needs, without restriction, and retrieve those resources easily (e.g., all resources associated with a clinical intervention or a workflow process). Group elements enable the user to describe the reasons for collocation and create multiple levels of specificity for this description. Elements that apply to these groups are ‘Description,’ ‘Date of Next Review,’ and ‘Visualization’.

The resource level denotes each distinctly named CDS resource. Elements around creation and maintenance are naturally applied here.

Critical points of interest, called events, were identified and parsed from each resource to create the resource part level. Events currently used are listed below (Table 1). ‘Logic’ and ‘Patient Type’ elements are captured around these events.

**Table 1. Events within clinical decision support rules.**

<table>
<thead>
<tr>
<th>Control orderables</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Modify</td>
</tr>
<tr>
<td></td>
<td>Discontinue</td>
</tr>
<tr>
<td>Control program flow</td>
<td>Execute</td>
</tr>
<tr>
<td></td>
<td>Load</td>
</tr>
<tr>
<td></td>
<td>Require specified exit checks</td>
</tr>
<tr>
<td></td>
<td>Exit without ordering</td>
</tr>
<tr>
<td>External notification</td>
<td>Send email</td>
</tr>
<tr>
<td></td>
<td>Send page</td>
</tr>
<tr>
<td></td>
<td>Log</td>
</tr>
<tr>
<td>User interaction</td>
<td>Display information/warning</td>
</tr>
<tr>
<td></td>
<td>Radio button</td>
</tr>
<tr>
<td></td>
<td>Checkbox</td>
</tr>
<tr>
<td></td>
<td>Dropdown</td>
</tr>
<tr>
<td></td>
<td>Button</td>
</tr>
<tr>
<td></td>
<td>Textbox</td>
</tr>
</tbody>
</table>

Each event is assigned a human-readable explanation of the logical condition that must be true for the event to execute. Additionally, each event has a unique identifier that is used to link events together. Finally, the value of the event is captured; for example, the specific file being loaded or the parameters of an order being modified are extracted from the source code.

Events are not tied to logical blocks, but are the points at which a resource affects the environment outside of itself, such as executing a program or interacting with a user or a different clinical system. It is common that several events will happen due to the same condition being true, or may be alternatives where a user will only be affected by one event from a set. For example, selecting to order Esmolol Infusion in a cardiac step-down unit requires electrophysiologist approval. Providing this approval places the order in the queue to be signed, logs the approving physician’s name, and emails a pharmacist. Each of these actions takes point in a different block of code: in fact, in two different files. Our interest in these operationally interconnected events is that they were all triggered by one action and that we may decide they should all happen together in the future system regardless of how many steps it currently takes to achieve that function. The loose structure of events therefore allows us to extract the clinical content and medical context of each event from the application’s execution requirements. In the VGR system, each CDS resource has several events, ranging from 3 to over 300, due to the needs of the HEO system for presenting alternatives to the clinician and providing real-time response without the delay of querying underlying databases after each selection. By defining clinically and operationally significant events rather than the blocks of code that exist now, future CDS implementations are not bound by past technological requirements.

In linking specific events to one another, only sequences have been preserved explicitly. Sequence is captured by human coders by expanding on all possible sets of events that generate all the data within that event’s condition and value. Alternatives in this sequence are allowed through a simple description logic that supports unions and intersections. Alternatives have been captured only implicitly, allowing future representations to reorder content or take pieces from resources if desired. Events are linked through backward chaining sequences, allowing the user to understand what prior data had to exist to make the current event possible. Together, all these individual data points
allow for robust visualization of highly connected networks that create a comprehensive picture (Figure 3) from initial states to ultimate possible outcomes.

**Figure 3.** Generic protocol visualization.
Red, blue, green, and black boxes represent events. Arrows indicate the direction of relationships. White boxes note conditions that must be fulfilled.

Controlled vocabularies come from taxonomies around decision support if possible, and local taxonomies if necessary. Specifically, ‘Category of Intervention,’ ‘Locus of Control,’ ‘Clinical Task,’ and ‘Method of Application’ come from taxonomies. For ‘Clinical Venue’ and ‘Patient Type,’ we had to create our own local taxonomy based on the values used in the local systems. Location data are inherently administrative and not able to be standardized across institutions, and grouping patients into similar types is also based entirely on local needs. To fully capture these elements, therefore, we started from the data definitions within our systems rather than an external source. This allows us to understand whether the current functionality can be exactly recreated or must be modified to work in a new system.

In any system allowing multiple users to create displays for different needs, there will be variability in interface design. Systematically capturing these techniques involved manual effort in both cultivating a controlled vocabulary and applying it. Display techniques captured in the ‘Category of Intervention’ element illustrate the evolution of a local controlled vocabulary. Each HTML page and display template overrides the common HEO interface and is manually coded. This has led to many different displays for clinicians that employ various tactics to present information effectively. The front-end tools taxonomy of Wright et al.\(^24\) provided a foundation, as front-end effects are often displayed in addition to executing – displaying a medication order sentence before placing that order, for example. Focusing only on information display, we removed several elements of the taxonomy that were purely functional and added new vocabulary elements for display techniques not presented in the taxonomy, such as ‘drug name/brand name display’. We also created controlled terms of general design techniques for successfully engaging
users with clinical systems from the human-computer interaction review of Horsky et al.29 By applying taxonomies of CDS design, we enable users to analyze different designs in a rigorous manner and reuse successful features, an idea presented as a potential benefit of taxonomies in Greenes.25 The exact moment in the clinical workflow that CDS is triggered is one aspect (‘Implementation Point’) captured in a taxonomy that allows users to compare differences between content initiated when an orderable is requested and content initiated after all orders are placed.

Scripting allows for the automation of some of the above tasks. Events were parsed from source code by exploiting the syntax of the underlying programming language and using regular expressions to extract and identify each event type. Based on the events present in a resource and the relationships to other resources found, the ‘Initializes,’ ‘Initialized By,’ ‘Related Orderable,’ and ‘Implementation Point’ elements were automatically inferred. System-level fields were automatically populated and passed to all resources within that system. Finally, elements that could be factually stated without ambiguity were also automated – ‘Revision,’ ‘Full Title,’ ‘Short Title,’ ‘Hover Description,’ ‘Filetype,’ and ‘Dependencies’.

Evidence is preserved in a separate workflow from the CDS it informs, as evidence around a concept can affect more than one resource. This conceptual grouping is an extension of the generic grouping feature. For example, clopidogrel metabolization is affected by genetic variation in the CYP2C19 gene. Furthermore, the evidence supporting differentiation in clopidogrel dosing by CYP2C19 variation is not simply one resource, but a collection of external resources, local evaluation of evidence quality and impact, and potentially even local practice constraints. Therefore, this concept of evidence-based practice regarding clopidogrel metabolization is the logical unit at which to maintain the evidence and local decisions on usefulness of the evidence. Attaching this concept to each resource that requires it (five resources in the current systems to perform additional decision support recommending alternative treatment for intermediate and hypo-metabolizers, and flag patients’ charts a predetermined color for those who have a recorded drug-genome interaction) allows us to periodically evaluate whether the evidence supporting this concept has changed and whether all our current decision support involving this concept is ideal.

Validity of all resources is ensured in three steps. First, the metadata for each resource is checked for accuracy by automated and manual processes. An automated check requires valid data for controlled fields, and accepts the links between specific events only if they are structurally possible between the two resources. Two manual checks, one by a CKM information scientist and one by a knowledge engineer, ensure that the human-readable logic and other event and resource-level fields are correct. Second, the resource is thoroughly tested in the production environment to determine it is still working correctly. Third, in order to determine what content is important to carry forward, a team of knowledge engineers is continuing to assess the clinical relevance of each resource by reviewing usage data, working closely with subject matter experts and system developers. At each of these points, documentation and resolution of any errors becomes part of ongoing content maintenance tasks. A companion document was created for each clinical system to provide the team with up to date information and examples of how each element should be correctly applied.

Discussion

Despite the many benefits of having CDS tightly coupled with the executing applications, this approach does require “the author to be familiar with not only the clinical domain but also with the syntax and features of the programming language.”28 For an outsider to a system to understand its decision support content required additional training in that system, even with clinical expertise and experience creating other decision support content. The VGRs were designed as a way to allow clinical experts to learn a simplified functional language rather than the programming language of HEO,4 but this benefit only applies to the single system. Ordersets are generated through a user-friendly tool that masks the programming complexities that enable each to work in the HEO application,29 but again this tool only assists in orderset creation and maintenance. We have successfully described content from two very distinct systems in a common format. Our project has shown it is possible to represent all decision support content in a consistent manner, within one single system. Each type of decision support being system-specific does require authors and editors to learn those respective tools, but content can be considered outside of the narrow bounds of the application through CS-KAAT.

In the guidelines modeling literature,16, 17 work is conducted from a textual representation to an executable one to allow different interested parties access points to the model. The CKM team worked in the opposite direction, from executables to semi-structured representation, to similarly create access points and enable flexibility to transition decision support to any number of potential systems. Multiple layers of abstraction allow end users to view content as complexly or simply as they desire. At the highest level, there is a basic description or a visual grouping. Below
In each of the above aspects, existing standards did not fully meet our needs. Our literature review and focus on discoverability highlighted several important descriptive facets that are not included in existing standards. Therefore, to fully meet our goals, it was important to adapt our schema with taxonomies extending standards. The guiding process for adding elements to the schema was to present a use case for each element to provide information that could not come from any existing element. This focus on institutional need led us to view each source as a potential contributor to the schema. With different needs, these extensions into literature-derived elements and local vocabularies will be different. For any institution interested in establishing a similar process, this has averaged 6.5 FTE equivalent effort. Given the diversity and complexity of the work, completing knowledge extraction and transformation for one named CDS resource has ranged from ten minutes to two days.

Separating the review processes for evidence and content allows flexibility in involving the appropriate parties for each step in the maintenance process. Knowledge management information scientists, who have extensive experience providing evidence, are well positioned to ensure that the latest evidence is provided to the content specialist and developers creating CDS resources. Then, clinicians can provide local context on best evidence to determine to create, modify, or retire CDS resources. This is anticipated to relieve some of the burden of engaging with CDS content for maintenance decisions. Consistent, externalized evidence provision also provides the institution the ability to more effectively integrate evidence with CDS at the point of care by making each relevant citation a focal point of the model, one that is purposefully saved and maintained. Finally, CS-KAAT provides the institution with a robust citation repository that is up to date, expertly curated, and has institution-wide scope.

CS-KAAT has made data dependencies traceable across systems. To illustrate the revelations that our tool has provided, we can follow an orderable from order creation in HEO to patient evaluation in StarPanel. A user initiates an order which is flagged to trigger a VGR when it is requested. From there, a series of VGRs eventually ends with the correct order being placed. After passing through an interface that links the different operational systems from the CPOE to the pharmacy system, data about the successful filling of this order are again sent through this generic interface engine to the medical record. An intermediate database duplicates information about the order of interest and distributes that information to the patient’s chart via an Indicator. The VGR and Indicator in this example are clearly related, but this information was obscured by the complex relationship between the two CDS resources. Because our schema has a facet for related orderables, the relationship is clearly seen at the semi-structured level and directly searchable. Additionally, the application of UMLS keywords to each of the records allows for files to be linked by concept, something that was also previously unavailable to users. Similarly, each facet gives users of CS-KAAT the powerful capability to understand connections and associations that previously only a few developers had.

Validation efforts enforced adherence to the schema. In each area where scripting was used for content extraction, users were able to manually override the harvested value. Periodic review of these overrides led to editing the scripts, making data collection more authoritative. We found that the compiler for the VGR local scripting language allowed slight variations to syntax, for example, and adjusted our extraction scripts to capture these. While these scripts are only useful in parsing knowledge from the current versions of local systems, this method both saved time for our team members and provides trusted content to end users. For data that could not be extracted automatically, validation by at least two people other than the documenter provides another level of assurance to end users that the content of CS-KAAT is trusted.

This was a significant effort that required many people to contribute different skill sets and to approach the work with a great deal of collaboration. The work at Partners and Intermountain Healthcare demonstrate that this is an extensive process, fraught with difficulty.\textsuperscript{31} It has been successful at VUMC through two key human features of our work: flexibility and communication. Our schema went through many early iterations, and as workflows changed, both the team and the CS-KAAT adapted to evolving needs. The CKM team provided support for each other and interchanged roles to assist in completing different portions of the work. Communication between teams and with stakeholders has allowed progress to be measured and created buy-in. For example, the user interface has been improved after first-time users from Knowledge Engineering provided usability feedback.

The separation of clinical content from application is critical to the long-term usefulness and maintenance of CDS.\textsuperscript{3} While there are knowledge editors for certain CDS types,\textsuperscript{30} these useful and highly used tools do not have institution-wide scope, a situation described as typical of most institutions today.\textsuperscript{31} CS-KAAT demonstrates an effective means
for separating clinical content from native applications, and further makes important steps towards an institution-wide knowledge editor through a unified acquisition and archival tool.

Future Work

The content within CS-KAAT will continue to be used to prioritize CDS for the new clinical systems, and as the new technical requirements of continued clinical needs are defined will be remixed to meet those requirements. The CKM team is focused on continued enhancements to the tool to enrich lifecycle support and enable retention of collaborative documentation for future use. As CS-KAAT has proven to be effective in the realization of our stated goals, the team will continue collaboration with Knowledge Engineering and others within the HealthIT group to incorporate CDS resources from other systems with the ultimate goal of representing all CDS content that affects care at VUMC. Further, the team is exploring various tactics for displaying evidence underlying CDS to clinicians at the point of care to enhance usefulness of decision support systems while retaining clinical user satisfaction.

Conclusion

VUMC has externalized decision support content from multiple sources in a consistent schema. Descriptive facets from decision support standards, as well as empirical taxonomies and controlled vocabularies, create many new ways for users to find and evaluate content. This understanding is supported through a focus on human-readable descriptions and multiple levels of representational complexity. In all, this flexible structure positions the medical center to effectively maintain the entirety of its decision support at the institutional level and continue to develop new methods of effectively supporting clinical care.

Acknowledgements

The authors would like to acknowledge the dedication of the CKM team members who have made this project a success: Mallory Blasingame and Jing Su. Additionally, we would like to thank all of the members of the Knowledge Engineering team, including Tina French, Shari Just, Sylinda Littlejohn, Janos Mathe, Michael McLemore, Lorraine Patterson, Debbie Preston, Audra Rosbeny, and Charlie Valdez, for their collaboration and contributions.

References

7. Raths D. Vanderbilt University Medical Center begins Epic journey. Healthcare Informatics. 2015.
Characterizing Physicians Practice Phenotype from Unstructured Electronic Health Records

Sanjoy Dey, PhD¹, Yajuan Wang, PhD¹, Roy J. Byrd¹, Kenney Ng PhD¹, Steven R. Steinhubl, MD², Christopher deFilippi, MD³, Walter F. Stewart, PhD⁴

¹IBM Research, T.J. Watson Research Center, Yorktown Heights, NY USA
²Geisinger Health System, Danville, PA USA and Scripps Health, San Diego, CA USA
³Inova Heart and Vascular Institute, Fairfax, VA USA
⁴Sutter Health Research, Walnut Creek, CA USA

Abstract

Clinical practice varies among physicians in ways that could lead to variation in what is documented in a patient’s electronic health records (EHR) and act as a source of bias to predictive model performance that is independent of patient health status. We used EHR encounter note data on 5,187 primary care patients 50 to 85 years of age selected for a separate case-control study covering 144 unique primary care physicians (PCPs). A validated text extractor tool was used to identify mentions of Framingham heart failure signs and symptoms (FHFSS) from the notes. Hierarchical clustering analyses were performed on the encounter note data for finding subgroups of PCPs with distinct FHFSS documentation behaviors. Three distinct PCP groups were identified that differed in the rate of documenting assertions and denials of mentions. Physician subgroup differences were not explained by patient disease burden, medication use, or other factors related to health.

Introduction

The widespread adoption of electronic health records (EHR) by US health care providers [1] is motivating a rapid growth in the use of predictive models to guide clinical decisions [2], to identify patients at high risk of future events (e.g., 30-day readmission) [3], and to detect disease early [4], among other applications. Copious longitudinal structured and unstructured data are captured by EHRs to characterize the patient’s demographic (e.g., age, sex, address), health and treatment status, diagnoses, lab test results, and medication orders. As much as 80% of the EHR data is thought to be in unstructured form [5]. To effectively use EHR data it is important to understand how the data comes to be.

Physicians are the dominant sources of the data captured in EHR. However, physicians vary substantially and systematically in their clinical practices [6] resulting in variation in what is ordered, diagnosed, and documented for each patient, in medication prescribing and in preferences for the intensity of practice. Such physician practice styles are not idiosyncratic. Rather, practice style is known to be directly or indirectly influenced by medical school training, regional practice standards, local practice standards, and performance incentives, among other factors. However, most of these clinical practice differences are independent of the underlying health status of the patient or other characteristics of patients such as demographics or prior genetic predisposition to the disease.

These “practice phenotype” differences can significantly influence the quality of both structured and unstructured data in the EHR and act as a source of potential bias for any downstream analysis of EHR data. To build accurate computational models, we need to detect and normalize for such variances in physician behavior. However, previous studies have largely focused on differences in physicians’ practices using structured EHR data to characterize diagnostic practice [6], regional practice pattern and standards [7], expertise, prior educational and training background of doctors [8], and the patient’s treatment plan [9]. Differences may be identifiable by a limited number of practice phenotypes. Prior studies indicate that patient utilization phenotypes can be identified from structured data that are strongly influenced by provider preferences [10]. But, documentation in unstructured data has not been examined for such patterns.

In this study, we aim to explore whether there are practice phenotypes that characterize differences among physicians in how information gets into the unstructured EHR data. To test the phenotype hypothesis we use physicians’ notes to determine whether there are practice phenotypes in the documentation of Framingham heart failure signs and symptoms (FHFSS). FHFSS are frequently documented in progress notes by PCPs, often years before HF diagnosis
We describe a systematic framework, based on a clustering approach, for characterizing the practice phenotype of PCPs using a large scale unstructured EHR data.

**Methods**

We focused on FHFSS to explore variations because physicians routinely document the presence or absence of the symptoms among older primary care patients independent of HF diagnosis. There are a number of challenges in extracting FHFSS from clinical notes that are actually related to practice phenotype. First, FHFSS are often documented in clinical notes much earlier than the clinical diagnosis of HF, which means that tracking the clinical notes longitudinally throughout the patient’s medical history is needed. Second, a physician’s practice might be affected by other confounding factors such as the patient’s age, sex, prior medical history, other co-morbidities and the physician’s expertise. Such confounding factors have to be removed in order to obtain an unbiased estimate of the actual practice phenotype defined by documentation behavior of FHFSS.

We carefully performed the study design and cohort selection to remove the effects of confounding factors during the feature extraction from clinical notes. A previously validated natural language processing (NLP) tool [12] was used for extracting FHFSS from clinical notes. For convenience, we used a large well characterized sample of primary care patients selected as controls for a prior nested case control study of heart failure [4]. We focused on the control group of patients to avoid any potential practice behaviors that might be a result of disease onset rather than actual practice phenotype of physicians. We were interested in broadly testing the hypothesis of physician documentation phenotypes in a representative sample of patients and not in a sample that was defined by any specific disease. Details of the patient sample and data source are summarized below along with the feature construction and analytic methods. This study was approved by the Geisinger Institutional Review Board.

**Study Design, Population, Setting, Source of Data**

Longitudinal EHR data were obtained on patients, 50 to 85 years of age, from the Geisinger Clinic, a multispecialty group practice that provides care to approximately 400,000 residents in central and northeastern Pennsylvania. EpicCare EHR was installed at Geisinger before 2001.

A total of 1684 incident HF cases among Geisinger primary care patients were identified over the time period from 2003 to 2010 [4, 13, 14]. Up to 10 eligible primary care clinic-, sex-, and age-matched (in 5-year age intervals) controls were selected for each incident HF case for a total of 13,525 Geisinger control patients. Primary care patients were
eligible as controls if they had no HF diagnosis up to the one year post-HF diagnosis of the corresponding HF case. Control subjects were required to have their first office encounter within one year of the incident HF patient’s first office visit and have ≥1 office encounters 30 days before or any time after the case’s HF diagnosis date to ensure similar duration of observations among cases and controls. Nine or 10 controls were identified for 49% of the Geisinger cases; 1.5% of Geisinger cases had only 1 to 2 controls.

The primary care physician (PCP) was the unit of analysis and patient data were nested within each physician. Patient data were assigned to the physician who was the designated primary care provider as documented in the EHR. Assignment of patients to a PCP is documented in an EHR structured field. Note that different encounters of the same patients might correspond to different PCPs, if that patient utilized care from multiple providers. In that case, we mapped each patient to the unique PCP who treated that patient for the longest period of time. Of the total 13,525 patients, 11,268 were explicitly assigned to a PCP.

The EHR data used for this study were selected from the time period in controls that was 12 to 36 months before the incident diagnosis of HF in the corresponding matched case. This time period prior to the diagnosis of HF of the matched case is denoted as the “observation window” (Figure 1). Patients were excluded if: 1) they did not have an encounter note in the 12-36 month observation window (this observation window retained the largest amount of patient data as described elsewhere [4]); 2) if the coverage duration period (i.e., total time span covering patient encounters with the PCP that are within the observation window) for the patient was less than or equal to 180 days; and 3) if a patient had a substantial number of documented chronic diseases during the observation window. This last step was used to minimize documentation variation among physicians that could be explained by patient’s comorbidities other than disease burden. We considered 1148 ICD-9 diagnosis codes coming from three different types such as Chronic Disease (200 codes), Cardiometabolic Chronic Disease (743 codes) and Chronic Episodic Disease (214 codes) with only 9 codes being common between the first two categories. Most of the patients have very low chronic conditions as shown in Figure 2. Therefore, we only kept patient with at most five (out of 1148 codes) chronic conditions to retain maximum number of samples. The first two criteria of the observation windows of 12 to 36 months and the minimum coverage of 6 months led to 6862 patients in total, while imposing all three criteria reduced the samples to 5,187 qualified control samples to 144 PCPs in total. The data used for analysis was confined to this subgroup of PCPs.

![Figure 2: Qualified patient distribution based on the frequency of chronic diseases from three different types: Chronic Disease (200 ICD-9 diagnosis codes), Cardiometabolic Chronic Disease (743 codes) and Chronic Episodic Disease (214). 284 out of 5351 patients (5%) had no chronic disease. Majority of patients (34%) had only two out of 1148 diagnosis codes.](image)

**Unstructured Data Extraction from Longitudinal EHR Patient Data**

We used physician encounter note data to extract mentions of FHFSS. The FHFSS were originally published in 1971 [15] and are often a focus of encounter documentation when physicians assess a patient’s cardiovascular health. A
hybrid natural language processing (NLP) tool called PredMED [12] was used to identify mentions of FHFSS within the notes and to label notes according to whether, at the encounter level, each FHFSS was asserted or denied. PredMED achieves an F-score of 0.910 for mention extraction and an F-score of 0.932 for encounter labeling. Among the 17 FHFSS described in [15], 15 were deemed clinically relevant based on prior work [4, 12] and given the known frequency of documentation (Table 1). PredMED extracted mentions of these 15 FHFSS, along with modifiers that indicated the assertion (presence) or denial (absence) of each condition [except Tachycardia and WeightLoss, for which only assertion was extracted]. The resulting 28 FHFSS were used as features, for further analysis of PCP documentation.

PredMED was applied to all progress notes that were created in the 24-month observation window. Summary statistics were derived for each patient by counting of number of FHFSS mentions in the observation window. Moreover, the counts were normalized for the total timespan of all encounters and the total number of encounters within the observation window (Figure 1). This resulted in a feature vector for each patient containing the fraction of total encounters per year that contains a FHFSS mention. In a next step, this patient level information was summarized into the physician level information. Specifically, the averages of all these FHFSS fractions of all the patients belonging to a particular physician was used to build the final feature vector for each physician.

**Statistical Analysis**

For analysis, each physician is represented as a feature vector of 28 FHFSS. We used unsupervised clustering to determine if there were natural groupings of the physicians. Clustering algorithms use a distance metric for computing the similarity between two sample points. For this analysis, we used Euclidean distance as the distance metric, since the features are represented as fractions of total encounters and normalized by the span of duration of care.

We explored several agglomerative hierarchical clustering analysis (HCA) techniques [16] to determine if natural groupings were identifiable. These techniques are easier to interpret than other clustering approaches such as agglomerative and density based clusterings [16]. For example, the hierarchy generated by HCA provides relationships among different samples and the obtained physician groups, which will be useful to characterize the physician groups. Among various versions of HCA techniques, we used the Ward based algorithm due to its inclination of finding globular clusters similar to K-Means, while preserving the hierarchy of the obtained clusters [16].

Next, we compared the physician groups against each other to characterize the different documentation behaviors of each cluster. We used several descriptive statistics and visualizations to facilitate interpretation. For example, principal component analysis (PCA) [17] was performed to visualize the identified clusters. Also, the mean documentation behaviors of each FHFSS was compared among the clusters using statistical t-tests.

### Table 1: 28 Framingham heart failure signs and symptoms (FHFSS) extracted from text notes using the PredMED text analysis tool. The mean, median and standard deviation of the fraction of encounters where each of these 28 features was reported by PCPs is shown.

<table>
<thead>
<tr>
<th>FHFSS Description</th>
<th>FHFSS Code</th>
<th>Assertion or Denial</th>
<th>Mean</th>
<th>Median</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilateral ankle edema</td>
<td>AnkleEdema (ANKED)</td>
<td>Assertion</td>
<td>0.157</td>
<td>0.120</td>
<td>0.142</td>
</tr>
<tr>
<td>Bilateral ankle edema</td>
<td>AnkleEdema (ANKED)</td>
<td>Denial</td>
<td>0.534</td>
<td>0.565</td>
<td>0.176</td>
</tr>
<tr>
<td>Acute pulmonary edema</td>
<td>APEdema (APED)</td>
<td>Assertion</td>
<td>0.006</td>
<td>0.000</td>
<td>0.032</td>
</tr>
<tr>
<td>Acute pulmonary edema</td>
<td>APEdema (APED)</td>
<td>Denial</td>
<td>0.006</td>
<td>0.000</td>
<td>0.032</td>
</tr>
<tr>
<td>Dyspnea on ordinary exertion</td>
<td>DOExertion (DOE)</td>
<td>Assertion</td>
<td>0.104</td>
<td>0.100</td>
<td>0.090</td>
</tr>
<tr>
<td>Dyspnea on ordinary exertion</td>
<td>DOExertion (DOE)</td>
<td>Denial</td>
<td>0.402</td>
<td>0.385</td>
<td>0.198</td>
</tr>
<tr>
<td>Hepatomegaly</td>
<td>Hepatomegaly (HEP)</td>
<td>Assertion</td>
<td>0.005</td>
<td>0.000</td>
<td>0.022</td>
</tr>
<tr>
<td>Hepatomegaly</td>
<td>Hepatomegaly (HEP)</td>
<td>Denial</td>
<td>0.365</td>
<td>0.330</td>
<td>0.208</td>
</tr>
<tr>
<td>Hepatojugular reflux</td>
<td>HJReflux (HJR)</td>
<td>Assertion</td>
<td>0.000</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>Hepatojugular reflux</td>
<td>HJReflux (HJR)</td>
<td>Denial</td>
<td>0.064</td>
<td>0.050</td>
<td>0.074</td>
</tr>
<tr>
<td>Central venous pressure &gt; 16 cm H2O</td>
<td>ICV Pressure (ICV)</td>
<td>Assertion</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Central venous pressure &gt; 16 cm H2O</td>
<td>ICV Pressure (ICV)</td>
<td>Denial</td>
<td>0.001</td>
<td>0.000</td>
<td>0.012</td>
</tr>
<tr>
<td>Neck vein distention</td>
<td>JVDistension (JVD)</td>
<td>Assertion</td>
<td>0.003</td>
<td>0.000</td>
<td>0.019</td>
</tr>
</tbody>
</table>
Results

Summary statistics were generated for each of the FHFSS. Table 1 contains the mean, median and standard deviation of the fraction of encounters where each of these 28 features was reported by PCPs. Overall, there are many more encounters with denials of the signs and symptoms than with assertions. A few of the signs and symptoms (e.g., AnkleEdema, DOExertion, Rales-denial, S3Gallop-denial) are relatively frequent. Several signs and symptoms (e.g., ICV Pressure, APEdema, WeightLoss) are very rare.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Denial</th>
<th>Assertion</th>
<th>Denial</th>
<th>0.239</th>
<th>0.180</th>
<th>0.193</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neck vein distention</td>
<td>JVDistension (JVD)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.012</td>
<td>0.000</td>
<td>0.033</td>
</tr>
<tr>
<td>Nocturnal cough</td>
<td>NightCough (NC)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.204</td>
<td>0.140</td>
<td>0.165</td>
</tr>
<tr>
<td>Nocturnal cough</td>
<td>NightCough (NC)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.006</td>
<td>0.000</td>
<td>0.024</td>
</tr>
<tr>
<td>Pleural effusion</td>
<td>PleuralEffusion (PLE)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.086</td>
<td>0.070</td>
<td>0.088</td>
</tr>
<tr>
<td>Paroxysmal nocturnal dyspnea</td>
<td>PNDyspnea (PND)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.028</td>
<td>0.000</td>
<td>0.054</td>
</tr>
<tr>
<td>Paroxysmal nocturnal dyspnea</td>
<td>PNDyspnea (PND)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.177</td>
<td>0.140</td>
<td>0.141</td>
</tr>
<tr>
<td>Rales</td>
<td>Rales (RALE)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.662</td>
<td>0.670</td>
<td>0.174</td>
</tr>
<tr>
<td>Rales</td>
<td>Rales (RALE)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.005</td>
<td>0.000</td>
<td>0.034</td>
</tr>
<tr>
<td>Radiographic cardiomegaly</td>
<td>RCardiomegaly (RC)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.014</td>
<td>0.065</td>
<td>0.152</td>
</tr>
<tr>
<td>S3 gallop</td>
<td>S3Gallop (S3G)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.007</td>
<td>0.000</td>
<td>0.025</td>
</tr>
<tr>
<td>Tachycardia (rate of ≥120 min⁻¹)</td>
<td>Tachycardia (TACH)</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.055</td>
<td>0.025</td>
<td>0.080</td>
</tr>
<tr>
<td>Weight loss of 4.5 kg in 5 days, in response to HF treatment</td>
<td>WeightLoss</td>
<td>Denial</td>
<td>Assertion</td>
<td>0.013</td>
<td>0.000</td>
<td>0.054</td>
</tr>
</tbody>
</table>

Figure 3: The dendrogram obtained from HAC with the partitioning for K=3 clusters based on Ward algorithm and Euclidean distance metric.
Table 2: Descriptive statistics for the three PCP clusters

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>p value (Group 1 vs 2)</th>
<th>p value (Group 2 vs 3)</th>
<th>p value (Group 1 vs 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of PCPs</td>
<td>63</td>
<td>61</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Patients</td>
<td>2860</td>
<td>1882</td>
<td>445</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. Number of Patients per PCP</td>
<td>45</td>
<td>31</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. Coverage Days</td>
<td>577</td>
<td>579</td>
<td>571</td>
<td>0.77</td>
<td>0.38</td>
<td>0.47</td>
</tr>
<tr>
<td>Avg. Age (Years)</td>
<td>70</td>
<td>70</td>
<td>69</td>
<td>0.99</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Female Gender (%)</td>
<td>51</td>
<td>51</td>
<td>57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. Count of Chronic Disease</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.30</td>
<td>0.31</td>
<td>0.68</td>
</tr>
<tr>
<td>Avg. Count of Cardio-metabolic Chronic Disease</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>0.78</td>
<td>0.26</td>
<td>0.18</td>
</tr>
<tr>
<td>Avg. Count of Chronic Episodic Disease</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.08</td>
<td>0.15</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Some basic descriptive statistics regarding the patients and physicians assigned to each of these three clusters are given in Table 2. Cluster 1 and cluster 2 were larger than cluster 3 in terms of number of patients and PCPs, but no significant differences were observed among the clusters in terms of coverage days, average age, gender, and comorbid chronic diseases. In this analysis, we used three categories of chronic diseases: “Chronic Disease”, “Cardiometabolic Chronic Disease” and “Chronic Episodic Disease”.

Figure 4: Scatter plot of the first and second PCA components show reasonable separation of the three PCP clusters.
Clustering Results as Groupings of Physician’s Behavior:

Based on HCA analyses, the PCPs were clustered into 3 groups with distinct documentation behaviors. We explored several values for the possible number of clusters (k), however k=3 produced more natural groupings obtained from the HCA dendrogram (Figure 3).

Analyzing the Behaviors of Physicians in Three Clusters:

We performed a principal component analysis (PCA) of the FHFSS features to assess the discriminatory power of the three clusters as shown in Figure 4. The first and second PCA components are represented by the horizontal and vertical axis, respectively. The three clusters have reasonable separation in the lower dimensional feature space.

We also analysed the average frequencies of FHFSS mentions in the three clusters. In Figure 5, we plot the contribution of each FHFSS to the each of the three clusters by taking the mean of count frequencies of all samples belonging to each cluster. Note that this figure is normalized by z-score for each FHFSS. All FHFSS except assertion of ICVPPressure varied among the three clusters to some extent. Overall, cluster 1 has the highest FHFSS counts whereas cluster 3 has the lowest. Cluster 2 contains medium counts of FHFSS.

![Figure 5. Mean FHFSS counts of each of the 28 FHFSS symptoms for the three clusters. (Darker shade means higher counts of FHFSS in the corresponding cluster).](image-url)
Contrasting the Behaviors of Physicians among Three Clusters:

We also looked for the specific practice variations of the physicians of each cluster by comparing them with other clusters in terms of the FHFSS frequencies. Figure 6 contains the individual fraction of visits for each of the 28 FHFSS. In addition, the right three columns of each of the two panels contain the pairwise comparison of the three clusters to assess whether there is a significant difference between the fractions of visits of the two clusters under consideration based on t-statistics. Only the FHFSS with p-value < 0.05 are marked in the last three columns. Group 1 PCPs (n=63) documented 10 out of 15 assertions, and 11 out of 13 denials of FHFSS significantly more frequently than Group 3 (n=20); while Group 2 PCPs (n=61) have significantly more frequent denial documentation behaviors than the other two (see Figure 6)

Discussion and Conclusion

EHR data contains information about both patient’s health characteristics such as the histopathological factors, demographics, treatment history and environmental effects as well as physician’s behavior such as treatment plans and orders and documentation behaviors of patient’s signs and symptoms. The availability of large-scale multi-source health data presents new opportunities and challenges for research that aims to effectively use these data to discover new knowledge to improve current health-care systems [18]. Such useful knowledge will not only help in personalizing healthcare for each patient with more accurate diagnoses, treatments and prevention plans, but also help reduce the unsustainable growth in healthcare cost.

Figure 6: Comparisons of mean documentation frequencies of FHFSS assertions and denials in 3 PCP groups clustered by HCA. The symbols indicate significant differences in pairwise comparisons by t-test (p<0.05). The horizontal axis shows the documentation frequency defined as the percentage of office visit encounters with FHFSS assertions/denials during the 2-year period. The vertical axis shows the assertion and denial FHFSS measure labels.
Finding patterns for a particular disease requires the secondary analysis of the EHR data collected from multiple sources. The recent growth of machine learning and data mining techniques offer a great help in analyzing large-scale healthcare data with new possibilities of developing predictive modeling for early detection of disease. However, traditional data mining and machine learning techniques often cannot be applied directly to EHR data because they are collected retrospectively in time and therefore can contain a lot of underlying bias and noise factors completely unrelated to disease burden [19, 20]. Unlike prospective studies such as randomized control trials (RCTs) which are designed to avoid such sources of noise and experimental biases, EHR data requires more careful analysis strategies to remove the effect of such noise and biases [21].

This study investigated physician behaviors as a source of bias in EHR data and a potential source of confounders for predictive modeling. PCPs were characterized by their documentation profiles of FHFSS. Distinct groups were identified for each of the profiles using hierarchical clustering analysis. Significant differences among the physicians’ practice in the three clusters were observed, which were associated with different documentation behaviors of FHFSS. Most (27 out of 28) of the FHFSS varied among these three clusters.

In terms of future work, we plan to investigate how to incorporate physician behaviors into predictive modeling and to quantify how much value (in terms of prediction performance) is added by eliminating this confounding factor in early detection of heart failure. Another interesting direction will be to further analyze the obtained groups of PCPs for finding their relationships with other potential causal factors such as physicians’ expertise, their training, and the geographic variations of healthcare providers as well as any other characteristics of patients corresponding to each group of PCPs.

Acknowledgement

This study was supported by the National Institute of Health (NIH grant No. R01HL116832). We are grateful to Zahra Daar, Heather Law, Elise Blaese, Harry Stavropoulos and Satish Mudiganti for a variety of contributions to this work, including project coordination, data preparation and database management.

References

5. ICTs, O., the Health Sector–Towards Smarter Health and Wellness Models. 2013, OECD Publishing.


Understanding patient satisfaction with received healthcare services: A natural language processing approach

Kristina Doing-Harris, PhD1, Danielle L. Mowery, PhD3, Chrissy Daniels, MS2, Wendy W. Chapman, PhD3, Mike Conway, PhD3
1Westminster College, Salt Lake City, UT;
2Director of Strategic Initiatives, University of Utah, Salt Lake City, UT;
3Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Important information is encoded in free-text patient comments. We determine the most common topics in patient comments, design automatic topic classifiers, identify comments’ sentiment, and find new topics in negative comments. Our annotation scheme consisted of 28 topics, with positive and negative sentiment. Within those 28 topics, the seven most frequent accounted for 63% of annotations. For automated topic classification, we developed vocabulary-based and Naïve Bayes’ classifiers. For sentiment analysis, another Naïve Bayes’ classifier was used. Finally, we used topic modeling to search for unexpected topics within negative comments. The seven most common topics were appointment access, appointment wait, empathy, explanation, friendliness, practice environment, and overall experience. The best F-measures from our classifier were 0.52(NB), 0.57(NB), 0.36(Vocab), 0.74(NB), 0.40(NB), and 0.44(Vocab), respectively. F-scores ranged from 0.16 to 0.74. The sentiment classification F-score was 0.84. Negative comment topic modeling revealed complaints about appointment access, appointment wait, and time spent with physician.

Introduction

Patient satisfaction ratings can be a good indicator of clinical effectiveness and patient safety1. However, the free-text comment fields, which are filled out in nearly 50% of patient surveys, are underutilized2. The Center for Medicare and Medicaid Services (CMS) and Agency for Healthcare Research and Quality (AHRQ) developed a national standard for reporting patient satisfaction called the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS). Siegrist et al.2 report that hospital participation in HCAHPS is very high. In fact, the number of surveys collected is higher than other customer satisfaction surveys. The level of patient response indicates that there is an enormous amount of free-text information available. In 2014, the University of Utah collected 105,000 free-text patient satisfaction comments. In order to understand the causes for patient dissatisfaction, quality improvement abstractors review every patient-generated comment, which is both labor-intensive and expensive. Large-scale automated or semi-automated review would save time and money. It would also facilitate scaling the analysis of these comments for benchmarking over time.

Analyzing Free-text comments

The information available in free-text comments has been identified using qualitative methods1,3,4. Lopez, et al.3 developed a complex taxonomy of patient comments (Table 1). It includes global themes of overall excellence, negative sentiment, and professionalism. They also identified specific factors, for example interpersonal manner, technical competence, and system issues. Doyle, et al.1 echoed Lopez’ et al.’s topic categories when they identify search terms for a meta-analysis of patient experience. They divide the terms into relational aspects and functional aspects. Relational aspects are equivalent to interpersonal manner. They include emotional and psychological support, patient-centered decisions, clear information, and transparency. Functional aspects are equivalent to Lopez et al.’s professionalism, technical competence, and systems issues. Functional aspects include effective treatment, expertise, clean environment, and coordination of care.

In terms of sentiment analysis, using a qualitative methodology, Lopez et al.3 categorized 712 online reviews of physicians from the websites ratemds.com and Yelp.com according to polarity and learned that most reviews were rated positive (63%) recommending a patient’s physician to others. Another study by Ellimoottil et al. reviewed physician rating sites for the scores for 500 urologists5. The free-text comments were classified...
as extremely positive, positive, neutral, negative and extremely negative. They found that most ratings (75%) were extremely positive, positive, or neutral. A meta-analysis of six physician rating websites also found that around 70% of ratings were positive.

Table 1: Lopez taxonomy of patient satisfaction themes

<table>
<thead>
<tr>
<th>Overall Excellence – Recommendation</th>
<th>Negative Sentiment – Intent not to Return</th>
<th>Professionalism</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPECIFIC FACTORS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal Manner</td>
<td>Technical Competence</td>
<td>System Issues</td>
</tr>
<tr>
<td>Empathic</td>
<td>Knowledgeable</td>
<td>Appointment Access</td>
</tr>
<tr>
<td>Friendly</td>
<td>Detailed</td>
<td>Appointment Wait Time</td>
</tr>
<tr>
<td>Helpful</td>
<td>Efficient</td>
<td>Practice Environment</td>
</tr>
<tr>
<td>Trustworthy</td>
<td>Clinical Skills</td>
<td>Practice Health IT</td>
</tr>
<tr>
<td>Time Spent During Appointment</td>
<td>Follow-up</td>
<td>Practice Location</td>
</tr>
<tr>
<td>Put at Ease</td>
<td>Referrals</td>
<td>Cost of Care</td>
</tr>
<tr>
<td>Listens</td>
<td>Perceived Poor Decision Making</td>
<td>Negative View of Healthcare</td>
</tr>
<tr>
<td>Explains</td>
<td>Perceived Successfulness of Treatment</td>
<td>Method of Physician Selection</td>
</tr>
<tr>
<td>Longevity of Relationship with Clinician</td>
<td>Complementary-Alternative Medicine</td>
<td></td>
</tr>
</tbody>
</table>

Natural Language Processing

Manually analyzing hundreds of thousands of free-text comments is difficult. Technology such as Natural language processing (NLP) can be used to make it more manageable. NLP has been used to classify comments by topics (topic classification), to encode the polarity of sentiment expressed within a comment i.e., positive or negative (sentiment analysis), and to determine if comments include unforeseen topics (topic modeling). Greaves et al. used topic classification to classify 6,412 free-text online comments about hospitals from the English National Health Service. They employed three topics – overall recommendation, cleanliness, and treatment with dignity. They created a Naïve Bayes multinomial classifier. It achieved F-measures of 0.89 (overall recommendation), 0.84 (cleanliness), and 0.85 (treatment with dignity).

There has been a lot of work in sentiment analysis both inside and outside of the patient satisfaction domain. In the SemEval-2016 task 4, sentiment analysis on Twitter, the winning team for the two-class (positive, negative) task used a combination of convolutional neural networks, topic modeling, and word embeddings generated via word2vec. They achieved an F-score of 0.80 and an accuracy of 0.86. Other investigators used keywords to indicate sentiment. These papers did not verify the accuracy of their vocabulary-based assessment. They relied on an accuracy measure of 75% for the dictionaries they employed. In a head-to-head comparison between commercial and non-commercial sentiment analysis tools for classifying healthcare survey data, Georgiou, et al. found that the WEKA implementation of Naïve Bayes’ performed the best, with a weighted F-measure of 0.81.

For topic modeling, the Brody et al. study, mentioned above, applied Latent Dirichlet Allocation (LDA) to 33,654 online reviews of 12,898 New York-based medical practitioners. Their model identified words associated with both specialty-independent themes (e.g., recommendation, manner, anecdotal, attention, scheduling) and specialty-specific themes (e.g., general practitioner: prescription and tests, dentist: costs, obstetrician/gynecologist: pregnancy). Maramba, et al. analyzed a free-text response from a post-consultation postal survey using a modified version of the English GP Patient Survey (GPPS) questionnaire. The question asked for any further comment. 3,462 individual comments were collected. They separated patient comments based on their overall rating of their experience on a 5-point Likert scale. Very satisfied and fairly satisfied were grouped as satisfied. Very dissatisfied and fairly dissatisfied were grouped as dissatisfied. The words “surgery”, “excellent”, “service”, “good”, and “helpful” were the five most distinctive words from satisfied patients, while the words “doctor”, “feel “, “appointment”, “rude”, and “symptoms” were the five most distinctive words in the comments from dissatisfied patients.

Our current work builds on these previous efforts by addressing patient satisfaction from Press Ganey patient satisfaction survey data gathered from a health care system rather than publicly accessible online reviews and NHS surveys. Additionally, we developed an NLP-powered classifier that combines vocabulary-based and machine learning-based methods to analyze both the topic and sentiment of each free-text comment. For the long-term goal of this project, we aim to leverage NLP methods to automatically analyze free-text fields in Press Ganey patient surveys at the University of Utah hospital system in order to streamline quality improvement efforts e.g., trending
historical patient experience data, helping direct future quality improvement efforts, and acquiring a better understanding of patient experiences as these relate to patient outcomes more generally. Our short-term goals are to determine common topics of patient satisfaction and dissatisfaction from free-text, patient survey comments, to create an NLP solution to automatically annotate comments with these topics, to analyze these comments for their polarity, and to identify sub-topics described within these comments to assist quality improvement efforts.

Methods

In this IRB-approved study (IRB_00081172), we obtained the 51,234 Press Ganey patient satisfaction responses from the University of Utah Health Care System (UUHCS) that were generated between January 1, 2014 and December 31, 2014. First, we developed a schema for characterizing topics from patient survey responses (Table 2) and validated our schema with an annotation study. Next, we trained two supervised classifiers (one a vocabulary-based classifier and one a Naive Bayes’ classifier) to automatically tag responses with topics from the schema. Then, we identified patient’s emotional valence toward these topics, using a separate trained classifier. Finally, we used LDA to cluster terms associated with negative experiences in an attempt to learn new topics.

Table 2. Annotation categories developed for this project.

<table>
<thead>
<tr>
<th>Advice_experience</th>
<th>Helpful_experience</th>
<th>Practice_environment_experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appointment_access_experience</td>
<td>Intent_not_to_return</td>
<td>Practice_family_friendliness_experience</td>
</tr>
<tr>
<td>Appointment_wait_experience</td>
<td>Intent_to_return</td>
<td>Professional_experience</td>
</tr>
<tr>
<td>Clinical_skill_experience</td>
<td>Knowledge_of_patient_experience</td>
<td>Recommendation_experience</td>
</tr>
<tr>
<td>Decision_making_experience</td>
<td>Knowledgeability_experience</td>
<td>Relationship_longevity_experience</td>
</tr>
<tr>
<td>Efficiency_experience</td>
<td>Listened_to_experience</td>
<td>Time_spent_experience</td>
</tr>
<tr>
<td>Empathy_experience</td>
<td>MyChart_experience</td>
<td>Treatment_success_experience</td>
</tr>
<tr>
<td>Explanation_experience</td>
<td>Overall_experience</td>
<td>Trustworthy_experience</td>
</tr>
<tr>
<td>Follow_up_experience</td>
<td>Patient_autonomy_experience</td>
<td></td>
</tr>
<tr>
<td>Friendly_experience</td>
<td>Percieved_bias_experience</td>
<td></td>
</tr>
</tbody>
</table>

Annotation Study

We developed an annotation schema for topics and sentiment recorded in free-text satisfaction responses. Starting with the taxonomy by Lopez et al.3 in Table 1, we added and removed categories in consultation with the Exceptional Patient Experience team (author CD) from UUHSC. In total, we created the 28 annotation categories listed in Table 2. Any group of words could be annotated separately with as many categories as were appropriate. Each annotation was also given a sentiment of positive, negative or neutral, an example is given in Figure 1.

Figure 1. An example of an annotated file.
A set of 300 documents was randomly selected from the dataset of patient survey free-text responses for manual annotation. Annotators A1 and A2 developed the annotation guidelines on batches of 100 documents until agreement was reached using consensus review. We considered inter-annotator agreement (IAA), F1-score, to have reached an acceptable level when the overall agreement, as reported by eHOST, was 0.74. A third annotator (A3 who has experience in the hospital quality domain) was trained in the application of the annotation schema. The annotations generated by the third annotator were compared to the adjudicated set. With our initial annotation scheme, we quickly found that sufficient IAA for all topics could not be reliably maintained. Therefore, we focused on the 7 most common (overall, appointment access, appointment wait, explanation, friendliness, practice environment, and empathy) with positive and negative sentiment, which represented 63% of all annotations.

Figure 2. Chart illustrating the flow of Patient feedback documents through our processing systems.

Topic classification

For topic classification, we developed both vocabulary-based and machine learning-based approaches. Our approach is illustrated in the top line of Figure 2. For the vocabulary-based method (our baseline approach), we used the annotated topics in the 300 adjudicated documents to generate a vocabulary. We gathered the text from each topic and used the Natural Language Tool Kit (NLTK) for Python* to tokenize the comments, remove non-alphabetic characters, stem (Snowball Stemmer) the resulting tokens, and remove stop words from each tokenized comment. We separated the results into n-grams: unigrams, e.g., “terrific,” and bigrams, e.g., “terrific service” (process A in Figure 2). The lists of n-grams were compared so that each n-gram appeared on only one list. If an n-gram was on more than one list, the list with the highest frequency of its occurrence got to retain it. The top five n-grams for each topic are listed in Table 3. From Table 3, we observed that the top n-grams associated with appointment access and appointment wait are associated with time. In contrast, practice environment n-grams are often associated with cleanliness and temperature. Empathy and friendliness n-grams describe feelings and service actions.

* http://www.nltk.org/
We compared two methods for classifying comments into each topic. First, we used a simple dictionary look up approach. For each unigram and bigram in a comment, we found the corresponding topic. All matched topics were retained because many comments had more than one hand annotation. Like all attempts at document classification there was a tradeoff between identifying the document category correctly (i.e., precision) and finding all the documents that belonged in that category (i.e., recall). We observed that allowing n-grams to appear in the lists of more than one topic increased recall at the cost of precision.

Table 3. The five most prevalent unigrams and bigrams for each topic category.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Type</th>
<th>N-gram feature set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>Unigrams</td>
<td>fantast, awsom, satisfi, absolut, fabul</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>good experi, great experi, excel experi, excel servic, far good</td>
</tr>
<tr>
<td>Appointment access</td>
<td>Unigrams</td>
<td>cancel, week, holiday, apart, saturday</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>schedul appoint, get appoint, abl get, get see, could get</td>
</tr>
<tr>
<td>Appointment wait</td>
<td>Unigrams</td>
<td>hr, end, period, realiz, paperwork</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>wait time, time minut, wait hour, long wait, exam room</td>
</tr>
<tr>
<td>Explanation</td>
<td>Unigrams</td>
<td>detail, futur, describ, bring, comdit</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>answer question, explain everyth, explain thing, explain would, happen futur</td>
</tr>
<tr>
<td>Friendliness</td>
<td>Unigrams</td>
<td>courteous, polit, interact, courtesi, paper</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>staff friend, alway friend, nurs friend, realli nice, feel like</td>
</tr>
<tr>
<td>Empathy</td>
<td>Unigrams</td>
<td>compassion, sensit, respect, encourag, situat</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>show concern, realli care, made feel, feel like, wait time</td>
</tr>
<tr>
<td>Practice environment</td>
<td>Unigrams</td>
<td>clean, wash, confirm, equip, thermomet</td>
</tr>
<tr>
<td></td>
<td>Bigrams</td>
<td>wash hand, alway clean, wait area, thermomet probe, hot drink</td>
</tr>
</tbody>
</table>

We first thought to increase precision by creating a dictionary listing for n-grams, which indicate the comment does not contain any of the topics of interest. By creating a list for comments that were “not annotated,” meaning that they did not contain any of the topics we were interested in. However, looking at our low recall scores, further stripping n-grams would be counter productive because it would cause more comments to be missed. So we trained a Naïve Bayes (NB) classifier using the 300 adjudicated documents to classify comments as “topic containing” or “no topic” (process B in Figure 2) before we used the vocabulary-based system and learned n-grams (keywords) on the “topic containing” comments. Otherwise, the system could always find at least one topic matched word in every comment leading to very poor precision scores.

For comparison against the vocabulary-based approach, we trained a machine learning approach leveraging the Naïve Bayes (NB) algorithm. We tested decision tree and SVM models as well, but we found the best performance with Naïve Bayes’. We used the NLTK (i.e., featx) methods to reduce the text to lowercase, stem words to their lemma, group the stemmed words into n-grams, and convert the n-grams into feature vectors. We used the NLTK NB classifier, with default settings. We divided the data set into training/test data sets (75%/25%), one binarized set per annotated topic, each set was balanced between “topic containing” and “no topic” comments (i.e., positive and negative examples of the class). This process is not pictured in Figure 2, in order to retain readability. We applied the trained algorithm to the blind test set. We report the results on the 25% test data compared to the vocabulary approach for the seven most common topics. Performance is reported as precision, recall and F-measure.

Sentiment Analysis

Using the same training and test sets and encoded n-gram features from the topic classification, we developed a classifier to categorize a comment based on its sentiment. In line with the head-to-head comparison performed by Georgiou, et al. and our results with the “topic containing”/”no topic” classification, we chose to again use Naïve Bayes’. Specifically, we trained the NLTK NB approach to classify comments based on sentiment categories of positive or negative (process C in Figure 2). We report the performance for each sentiment category on the test set.
We then ran both the trained and tested NB topic classifier and NB sentiment classifier on the remaining roughly 50,000 patient satisfaction comments. We report the distribution of positive and negative comments by topic class.

**Topic Modeling**

To complement the topics annotated through manual review, we completed an unsupervised topic modeling study. We first classified the full 51,234 comments with sentiment categories using the NB sentiment classifier (process C in Figure 2, repeated in the figure for clarity). For all comments classified as negative, we provided the unigrams and bigrams to an LDA algorithm with a preset maximum of 30 topics using the gensim package* We report the n-grams associated with 10 of the 30 topics learned by the algorithm and if a topic suggests one of our seven most common topics, we also provide a topic label (process D in Figure 2).

**Results**

**Annotation Study**

In total, we annotated 1,374 documents consisting of 2,021 annotations. All three annotators annotated the same 300 documents in 100 document sets. Figure 3 illustrates how IAA changed across the three annotators and 3 document sets. In general, there was little improvement between sets 1 and 2. Some categories’ performance even decreased indicating that new comments were difficult to categorize. Between sets 2 and 3 there is a general increase in IAA indicating that annotators had reached a common understanding. Empathy and practice environment categories, however, still demonstrate a lack of agreement.

Annotators A1 and A2 discussed categories and re-annotated sets 1 and 2 until they reached an understanding. On set 3 their agreement was 0.74 overall as reported by eHOST. At this point the most frequent topics (overall, appointment access, appointment wait, explanation, and friendliness) had high agreement levels of (0.83, 0.91, 0.79, 1.00, 0.76, respectively). Empathy (0.47) and practice environment (0.44) were much lower.

The same procedure of discussion and re-annotation of the first 2 sets was repeated with A3. His overall agreement with the adjudicated document set from A1 and A2 was reported by eHOST to be 0.73 overall. Agreement between A3 and the adjudicated documents for the most frequent categories was 0.77, 0.83, 0.77, 0.86, and 0.77, respectively. Empathy (0.67) and practice environment (0.60) demonstrated lower agreement, although not as low as the agreement between A1 and A2.

Taken across the 2 sets of IAA scores, the categories with the most agreement are explanation (1.00, 0.86), appointment access (0.91, 0.83), overall (0.83, 0.77), appointment wait (0.79, 0.77), friendliness (0.76, 0.77), empathy (0.47, 0.67) and practice environment (0.44, 0.60). Agreement with A3 is tied for the middle three. So the score between A1 and A2 is used to decide order.

*https://radimrehurek.com/gensim/models/ldamodel.html
unigram could not appear in more than one list, the number of unique n-grams is the number of words that were available to identify topics. 58 bigrams (20 of them unique) were found for overall experience. Figure 4 shows the counts of unigrams and bigrams, with the number of unique in each case, across the seven most common topics. Practice environment, appointment access, and appointment wait generated the most n-grams and the most unique n-grams (Figure 4). Looking ahead at the classification results in Table 3, there is no obvious relationship between the number of unique n-grams and performance. Explanation had the fewest unique n-grams and the highest performance. Overall experience had the most unique n-grams was directly in the middle of the performance scores.

Figure 4. Total n-gram counts and unique n-gram counts for each topic. Unigrams = light color (bottom); Bigrams = dark color (top).

The vocabulary was extracted from the 300 adjudicated documents and tested on the remaining 1,074. Table 4 illustrates the results of the topic classification for the top seven topics. NB classification for explanation was good (0.74). Performance for the remaining 6 topics were fair, ranging from 0.36 – 0.57. Generally, precision was higher than recall for the vocabulary matching approach. In choosing the vocabulary, we opted for higher precision. However, this strategy was not successful for the more difficult categories of practice environment and empathy. These were the only topics for which the vocabulary matching outperformed the machine learning approach of applying NB. The precision for the NB overall classifier was high, but recall was lower. The topic classification results echo the IAA results, demonstrating the difficulty of the task. Explanation was the best performer in both places. Empathy and practice environment are at the bottom. Friendliness and overall do not follow this trend.

Table 4. Topic classification for n=1,074 annotations. Classifiers are listed by their highest F-measure. IAA indicates the rank order of the topic IAA scores. Bold indicates highest performance between approaches.

<table>
<thead>
<tr>
<th>Topic</th>
<th>IAA</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Vocabulary</td>
<td>Naïve</td>
<td></td>
</tr>
<tr>
<td>Explanation</td>
<td>1</td>
<td>0.42</td>
<td>0.90</td>
<td>0.38</td>
</tr>
<tr>
<td>Appointment wait</td>
<td>4</td>
<td>0.31</td>
<td>0.54</td>
<td>0.32</td>
</tr>
<tr>
<td>Appointment access</td>
<td>2</td>
<td>0.45</td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td>Practice environment</td>
<td>7</td>
<td>0.38</td>
<td>0.24</td>
<td>0.61</td>
</tr>
<tr>
<td>Overall</td>
<td>3</td>
<td>0.66</td>
<td>0.85</td>
<td>0.30</td>
</tr>
<tr>
<td>Friendliness</td>
<td>5</td>
<td>0.50</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>Empathy</td>
<td>6</td>
<td>0.41</td>
<td>0.18</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Sentiment Analysis

Separate vocabulary and NB classifiers were developed to analyze the sentiment of a comment. These classifiers used all 1,374 documents split randomly into 75% for training and 25% for testing. The algorithm performed with a precision of 0.90 and recall of 0.80 for positive sentiment, and a precision of 0.79 and recall of 0.90 for negative sentiment. The overall F-score of the system was 0.84.

Combining Topic and Sentiment

We applied the two NB classifiers trained on 1,374 documents. At this point we trained the model on 100% of the documents because we used the trained model to annotate the 49,860 un-annotated patient satisfaction comments.
The classifiers tagged the comments with 73,801 annotations. Overall patients have positive experiences. Specifically, empathy, friendliness, and explanation are more often positive experiences; in contrast, appointment wait, appointment access, and practice environment are more often negative experiences. The breakdown of polarities for the top 7 topic categories is illustrated in Figure 5.

**Figure 5.** Distribution of positive and negative comments according to topic.

**Topic Modeling**

We applied a topic modeling algorithm to patient satisfaction comments from 2014 that had been automatically flagged as negative by our sentiment analysis tool to provide a snapshot of negative topics (Table 5). Topic models can be difficult to interpret\(^\text{14}\), but it can be seen that most of the themes identified refer to discontent regarding the topics we named appointment wait and appointment access.

**Table 5.** Topic words associated with each label from 10 of the 30 topics learned.

<table>
<thead>
<tr>
<th>Topic #</th>
<th>Topic words</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>back, called, minutes, waited, come, doctor, behind, running, told, would, see, receptionist</td>
<td>appointment wait</td>
</tr>
<tr>
<td>2</td>
<td>time, patient, help, spent, bit, short, kind, enough, doctor, needed, amount, year</td>
<td>time spent with provider</td>
</tr>
<tr>
<td>3</td>
<td>one, another, first, saw, seeing, weeks, physician, eye, doctor, quickly, contact, would</td>
<td>appointment wait</td>
</tr>
<tr>
<td>4</td>
<td>wait, waiting, room, time, exam, delays, delay, minutes, informed, taken, times, area</td>
<td>appointment wait</td>
</tr>
<tr>
<td>5</td>
<td>felt, almost, like, pain, something, found, bad, brought, rushed, although, believe, scheduler</td>
<td>appointment access</td>
</tr>
<tr>
<td>6</td>
<td>clinic, call, office, phone, appointment, need, days, would, back, someone, got, center</td>
<td>appointment access</td>
</tr>
<tr>
<td>7</td>
<td>much, hours, two, wish, appreciate, seems, clinic, would, hard, question, convenient, sick</td>
<td>appointment wait</td>
</tr>
<tr>
<td>8</td>
<td>get, able, schedule, appointment, appt, always, lot, done, around, see, talk, trying</td>
<td>appointment access</td>
</tr>
<tr>
<td>9</td>
<td>appointment, visit, seen, scheduled, time, even, early, day, arrived, right, actually, though</td>
<td>appointment access</td>
</tr>
<tr>
<td>10</td>
<td>made, appointment, appointments, since, last, seemed, experience, appointment access available, month, years, months, helped</td>
<td>appointment access</td>
</tr>
</tbody>
</table>

**Discussion**
In our topic classification study, we found that the common topics of patient satisfaction and dissatisfaction from free-text, patient survey comments were appointment access, appointment wait, empathy, explanation, friendliness, overall experience and practice environment. In just over half of the cases (four out of seven) the NB classifier was more successful than the vocabulary-based system in classifying documents as belonging to the topic in question. Investigating comments for their sentiment, using an NB classifier over the entire document set, we found 70% to be positive and 30% to be negative. This breakdown in sentiment was also found in our manual annotations where 71% of the top 7 annotations and were positive and 75% of all annotations were positive. Using topic modeling to determine if there were topics we had not considered within the negative comments, we found that the topics described within the negative comments, reflect the common topics of appointment access and appointment wait.

Our initial plan was to create a topic classification system that would reflect the full complexity of the topics found using qualitative methods. We created an extensive taxonomy of topics, which could be annotated with reasonable overall agreement. However, IAA was only acceptable for five of the seven most common categories. Empathy and practice environment, although they were commonly used were not often agreed upon. In the case of empathy, it may be due to the difficulty in distinguishing between empathy and related topics, like friendliness. The relatively low agreement for the practice environment category is perhaps due to conflating overall negative experiences with issues relating to the physical environment.

The ability to create a high performing topic classifier may have been affected by the only fair agreement on annotations. The classifier with the best overall F-measure was explanation, which had the highest IAA. The worst performing classifier empathy was for the topic with one of the lowest agreement scores. The middle four NB classifiers did not show a relationship between F-measure and IAA rank. This pattern did not hold for the vocabulary-based systems. The best vocabulary system was for practice environment, which had the second lowest agreement rank. The vocabulary system with the lowest F-measure was for appointment wait, which had a middling agreement rank. We also found that as we added documents that had been annotated by only A3, the classifier performance decreased. It seems likely that A3 experienced drift in their annotations.

Low performance on topic classification, even with our restricted topic set, indicates that perhaps both approaches are unable to capture the semantic information in the patient comments. Performance may be improved by creating an ontology of the dictionary terms, a structure that reflects synonymy, hierarchical relationships between terms, and term function. For instance, “excellent” may be synonymous with “great”, in this context. “Really excellent” may be an intensified version of “excellent” and both “really excellent” and “excellent” can be used in combination with “environment” or “interactions.” The former would be associated with the topic practice environment, while the latter with friendliness. This kind of information may be encoded in convolutional neural networks, which have been successfully used in sentiment analysis.

By focusing on solely on those topics on which annotators can reach acceptable agreement, and conflating easily confused topics, we may be able to dramatically increase our classification performance. Perhaps combining the categories of empathy, friendliness and helpfulness (from our bigger list) we could create a topic that is easier to distinguish from the other topics. It is also possible that the number of annotations of each topic led to poor performance. However, the best performing classifier was for explanation, which had far fewer occurrences (92) than overall (587), which had one of the worst performing classifiers. Future work includes training on a larger set of data and using more semantic features e.g., encoding words from categories such as ‘Time, Affect, Positive Emotion, and Negative Emotion, etc. from the Linguistic Inquiry Word Count lexicon’.

The sentiment classifier performed very well, which is consistent with the larger training set due to a smaller number of classes. Overall system performance was above the threshold of 0.80. It performed within the range of Greaves, et al.\textsuperscript{7,15,16} three-topic classifications e.g., 0.84 vs. 0.89, 0.84, and 0.85. Applying this classifier to our large dataset, we found 70% had positive sentiment compared to Lopez, et al.\textsuperscript{3} who found 63% of online reviews had positive sentiment.

Good sentiment analysis performance allowed us to select the comments with negative sentiment for topic modeling. Our topic modeling results echo our topic classification results in that the predictive topic words indicate two of our most common topics appointment access and appointment wait. These two are also the topics with the most comments with negative sentiment. They are the only topics with more than 50 comments with negative sentiment across the annotated dataset. Brody et al.\textsuperscript{4} also found scheduling to be a common review theme. We may have been more likely to find unexpected topics if we had looked at the least commonly occurring words in the negative comments. However, finding uncommon, unexpected topics may not be useful for quality control.

\textsuperscript{*} http://liwc.wpengine.com/
As we have noted in this discussion, the main limitation of this work is the relatively poor performance of the topic classification and topic modeling systems. However, the problems encountered are common to NLP systems. In the case of topic classification, the more topics you have the more difficult the task becomes for both humans and machines. For topic modeling, the system’s performance was not a problem. The results simply show that there are no unexpected topics found in the most frequent negative comments.

Conclusion

This study is a promising start to creating a method of analyzing free-text comments in patient satisfaction surveys that will help streamline patient satisfaction efforts e.g., trending historical patient experience data, helping direct future quality efforts, and acquiring a better understanding of patient experiences more generally. We are actively improving topic classification algorithm performance, expanding the granularity of sentiment classes to address strength of emotional valence (strong positive, weak positive, neutral, weak negative, and strong negative sentiment) and implementing monthly reports with statistics and visualizations to streamline patient satisfaction improvement efforts at UUHSC.

Acknowledgements

We would like to thank the anonymous reviewers for valuable comments. This work was partly funded by the University of Utah Healthcare System Hospital Project funds.

References

Dynamic Multicore Processing for Pandemic Influenza Simulation

Henrik Eriksson, PhD1, Toomas Timpka, MD, PhD1,2, Armin Spreco2, Örjan Dahlström, PhD2, Magnus Strömgren, PhD3, Einar Holm, PhD3

1 Dept. of Comp. and Inform. Sci., Linköping University, Sweden; 2 Dept. of Medical and Health Sci., Linköping University, Sweden; 3 Dept. of Social and Economic Geography, Umeå University, Umeå, Sweden

Abstract

Pandemic simulation is a useful tool for analyzing outbreaks and exploring the impact of variations in disease, population, and intervention models. Unfortunately, this type of simulation can be quite time-consuming especially for large models and significant outbreaks, which makes it difficult to run the simulations interactively and to use simulation for decision support during ongoing outbreaks. Improved run-time performance enables new applications of pandemic simulations, and can potentially allow decision makers to explore different scenarios and intervention effects.

Parallelization of infection-probability calculations and multicore architectures can take advantage of modern processors to achieve significant run-time performance improvements. However, because of the varying computational load during each simulation run, which originates from the changing number of infectious persons during the outbreak, it is not useful to use the same multicore setup during the simulation run. The best performance can be achieved by dynamically changing the use of the available processor cores to balance the overhead of multithreading with the performance gains of parallelization.

Introduction

Pandemic simulation can provide insights in outbreak dynamics and offer valuable information about the impact of different interventions on simulation outcome [1] and researchers have developed several frameworks to support simulation [2,3,4]. However, there are several challenges in pandemic simulation, such as community and disease modeling and development of simulation engines with sufficient performance. Run-time performance is a serious bottleneck because extensive turnaround times make it difficult for public-health analysts to run larger simulations and explore alternative scenarios. Improved run-time performance can create new possibilities for public-health researchers and decision makers to take advantage of pandemic simulation in novel ways, such as scaling to larger models and using fine-grained population and transportation data. In addition, improved performance can ultimately enable interactive simulations where the results change as users vary input parameters and simulation settings.

Discrete-event simulators based on mixing groups model the community as a set of locations where people meet and where transmission may occur. These simulators calculate the probability for infectious members of the mixing group to infect other susceptible individuals. The simulations are probabilistic in the sense that they use random-number generators to determine if individuals get infected, which means that the outcome varies among different simulation runs and that it may necessary to repeat the simulation several times to obtain sufficiently stable results.

The number of floating-point operations required for computing probabilities of infection is often a serious performance bottleneck. Run-time performance is important for scaling the problem to larger populations, running several different scenarios and stepping model parameters, and providing time-critical decision support. The run-time performance challenge is exacerbated by the requirement to repeat the same simulation several times, for instance 10 or 100 times.

The run-time performance is to some extent unpredictable because it depends on the evolving outbreak simulation. When the number of infected persons in the simulated outbreak increases, the number of possible interactions will increase as well, which requires additional computations (mainly floating point operations) to determine the probability of infections. Smaller outbreaks will run faster whereas larger outbreaks take longer time, and they are especially slow at the peak of the outbreak. The use of large mixing groups, such as models of large neighborhoods
and the entire simulated population, will increase the number of interactions and, thus, the run time required for simulation.

We have previously addressed the modeling challenge and introduced the use of ontologies for modeling of influenza outbreak scenarios [5] and for documenting the assumptions made during the modeling process [6]. The aim of this study, however, is to explore parallelization for improving run-time performance in pandemic simulation. Here, we describe an approach to use OpenMP [7] compiler directives for annotating a preexisting simulator and evaluate the impact of using multiple cores for computing probabilities in a pandemic simulator based on mixing groups. Furthermore, we introduce a method for dynamic switching between single and multicore mode based on a threshold for the number of infectious individuals. Finally, we discuss potential approaches for further run-time performance improvements.

Background

It is possible to parallelize pandemic simulation in different ways. If several alternative and independent scenarios are to be evaluated, one approach is to divide the work among several computing nodes, such as nodes as part of a cloud-based platform [8]. For example, simulation studies of outbreaks in communities where alternative intervention strategies are to be compared could run each of the cases on different machines. The disadvantage of this approach is that it only scales to the number of scenarios to run simultaneously (i.e., at most one node per scenario).

For stochastic simulations, where it is necessary to perform multiple runs of the same simulation with the same parameters, it is possible to divide the work of running such iterations among an array of independent computing nodes. This approach scales better than parallelizing scenarios because of the typical number of iterations. For example, 100 iterations could potentially be divided among 100 worker nodes. In this method, it is necessary to ensure that any random-number generator is properly seeded to make the iterations unique and independent. Because of the distribution of iterations over several nodes, there will naturally also be some additional work of collecting and merging the results from the different iterations.

Running several simulations in different operating-system processes on multicore CPUs, rather than on independent machines, may seem like a straightforward way of improving performance. However, this approach does not scale as well because of resources shared among the cores, such as RAM, and limitations in cache memory and memory bandwidth. For example, running four instances of the simulation program on a four-core processor does not result in four times the performance. To fully use modern multicore processors, it is necessary to develop threaded simulators that take advantage of the cores specifically (e.g., by multithreading).

Open Multi-Processing (OpenMP) is a portable programming model for shared-memory multiprocessing [7]. OpenMP provides compiler directives (#pragmas) and libraries for adding parallelism to the programming languages C, C++, and Fortran. OpenMP is designed to give programmers a high-level and an easy-to-use way of adding parallelism to new and preexisting programs. For example, OpenMP takes care of forking and running multiple threads (e.g., on a multicore platform) as well as collecting results from multiple threads running in parallel. In the context of pandemic simulators, one of the main advantages of OpenMP is that it allows for parallelization with minimal changes to the original simulator code and the core algorithms, which is beneficial for simulator transparency and maintainability.

Methods

The overall evaluation method is to run repeated simulations with different settings and compare the run-time performance. The performance evaluation used an influenza outbreak scenario as a benchmark for testing the impact of multicore processing. First, we extended the pandemic simulator with compiler directives for OpenMP. In particular, the central loop for calculating infection probabilities for each step in the simulation was augmented with a loop-reduction annotation to parallelize the loop over a number of cores (which was stepped from 1 to 32).

The experiments ran on an HP Z620 workstation equipped with dual Intel Xeon E5-2687W v2 eight-core processors running at 3.4 GHz and 96 GB DDR3-1866 ECC RAM. We specifically selected and installed 3.4 GHz eight-core processors because they provide a reasonable tradeoff between clock frequency and the number of cores deemed suitable for the mixed workload of pandemic simulation. (There are certainly alternative processors with more cores...
but lower clock frequency available, which would have penalized single-thread operation and resulted in less performance for large parts of the simulation.)

The experiment machine ran x86_64 GNU/Linux Debian. Processor hyperthreading was enabled, which means that the total number of cores presented to the operating system was 32. The simulator software was implemented in C++ and was compiled on the experiment machine using gcc with -O2 optimization and -fopenmp settings. All simulation runs were CPU bound rather than memory bound, which means that the amount of installed RAM had negligible effect on performance.

The simulator was extended with additional command-line switches for controlling the relevant settings (such as the number of thread and the threshold for invoking threading) and a shell script was used to run the experiment series under various settings and stepping parameters. The results were stored automatically in a database (SQLite3) and later extracted for plotting using Gnuplot.

Results

Figure 1 shows the benchmark simulation run-time for different numbers of cores and for different thresholds for activating threading. As illustrated by this diagram, the execution times drop dramatically (from 580 s) when enabling the first few cores, but tend to increase again as additional cores are used (due to overhead). Also, note the slight jump in the execution times when going beyond 16 physical cores (and thus using hyperthreading). Furthermore, it is clear from these results that using a threshold for the number of infections persons improves overall performance. There is a slight advantage of using a threshold of 400 infectious individuals over a value of 100, but the difference is not that large. For no threshold, the fastest run time was 144 s using 8 cores, whereas a threshold of 100 achieved 89 s with 13 cores and a threshold of 400 achieved 82 s with 15 cores.

![Figure 1. Simulation run time versus number of cores for different thresholds for activating multithreading.](image-url)
Figure 2. Simulation run time versus number of cores for an intervention scenario where schools close at day 10 and later reopen at day 20.

Figure 3. Simulation run time versus number of cores for an intervention scenario where schools close at day 10 and later reopen at day 20.

Figure 2 shows the effect on the execution times for an alternative scenario with a somewhat different simulator load. Here, the schools close on day 10 and are kept closed until day 20 when they reopen [9]. This intervention will reduce the number of infected during that period. However, the total number of infected during the duration of the outbreak will be similar. For this scenario, the single-core execution time was 564 s, and the best performance with no threshold was 156 s using 9 cores. For a threshold of 100, the fastest run time was 98 s using 12 cores and for a threshold of 400 the fastest time was 91 s using 13 cores.
Likewise, Figure 3 shows the result of closing schools at day 20 and reopening them at day 30. In general, the run-time results from the school-intervention scenarios in Figure 2 and Figure 3 are quite similar, which illustrate that the run-time results apply in different scenarios that are relevant for public-health simulation tasks. For the scenario of closing schools at day 20 and reopening them at day 30, the single-core execution time was 539 s. Using no threshold, the best performance was 144 s with 7 cores. For a threshold of 100, the fastest run time was 93 s using 9 cores and for a threshold of 400 the fastest time was 87 s using 11 cores.

The results presented previously (Figure 1–Figure 3) assume that a single simulation process instance was running at the time. The question arise whether it is possible to better use the cores available in the hardware by limiting the number of cores used and instead use remaining cores for running another simulation process instance. Figure 4 shows the results from running two simulation jobs simultaneously on the same machine. With two simultaneous jobs, the single core execution time (i.e., for each job) was 583 s, and the best performance with no threshold was 143 s with 8 cores. For a threshold of 100, the fastest run time was 106 s using 8 cores and for a threshold of 400 the fastest time was 100 s using 8 cores. These results suggest that it is possible to achieve better throughput by using a limit of eight cores and run two simultaneous simulation jobs (e.g., two scenarios), at least in the context of the benchmark and the dual eight-core processor machine used in the experiment.

**Discussion**

Public health needs improved simulation methods. Advances in simulation performance can benefit the general public by improving public-health decision processes, for example by allowing needed interventions to start faster and avoiding pointless interventions, which burden the communities. We believe that it is essential to develop new methods for improving simulation run-time performance. Such performance gains can allow for interactive sessions where public-health officials and decision-makers explore different scenarios and get a better understanding for the dynamics of an ongoing outbreak. Furthermore, improved performance will make it possible to use interactive simulation as a training tool where the outcome of different decisions can be illustrated through the simulation results. A better understanding of different outbreak and intervention scenarios combined with lessons learned from desktop exercises can contribute to resilience and healthcare preparedness.
In this study, we used a method of taking advantage of modern multicore processors based in a single threshold (i.e., 100 or 400) for activating multithreading. In principle, it is possible to use multiple levels of multicore activation, for example by employing several threshold levels and to ramp up the number of cores gradually. We have found that this method can have some performance benefits, but in general a single threshold is sufficient for controlling the activation of parallelism (because of the overhead of initializing the threading mechanism). However, in situations where several simulation processes run on the same machine simultaneously, there may be an advantage of gradually stepping up and down the number of cores used for the purpose of freeing up resources for the other simulation processes.

There are other alternatives for further improving the simulation run-time performance. One approach is to cache intermediate results and reuse them later in the simulation. For example, it is possible to store the probability of someone getting infected in a certain mixing groups and to reuse this result for another person with similar properties (such as belonging to the same age interval and otherwise have similar properties according to the simulation model). When appropriate, this approach can reduce floating-point operations and improve performance significantly. The drawback of this approach, however, is that caching is only possible for relatively basic scenarios and that reuse of intermediate computations steps are inapplicable for certain types of interventions affecting transmission probabilities (e.g., lowered spreading probabilities for some contagious individuals due to shielding measures).

Another potential approach to improve simulation run-time performance is to take advantage of offloading to GPU devices. GPUs support a massive number of threads and simultaneous floating-point operations, but are simpler devices than CPUs. Offloading typically requires transfer of data to and from the GPU-card over the PCIe bus, which can become a bottleneck. It is possible to use OpenACC [10], which is a standard for compiler directives for CPU/GPU architectures, in the same ways as OpenMP to annotate source code for compilation to CUDA (a platform for parallel computing developed by NVIDIA) and eventually execution on GPUs. However, current compiler support for OpenACC is somewhat limited1. An alternative to OpenACC is to develop directly in CUDA. Because of the bandwidth limitations in transferring data to and from the GPU and the associated overhead, it is difficult to merely accelerate performance-critical regions in the code. A serious implementation in CUDA would require a large proportion of the simulation to take place on the GPU, which essentially means a complete simulator reimplementation for CUDA. Our preliminary tests show that just preparing and packaging the relevant simulation data for transfer to the GPU can easily take more time than simply performing the calculation on multicore processors. Nevertheless, we believe that offloading to GPU is a promising future option because of the potential for scaling simulations to much larger populations and for developing truly interactive simulations where public-health professionals and decision makers can manipulate interventions and see the results instantly.

**Conclusion**

Multicore programming can significantly improve run-time performance of pandemic simulation by allowing for parallelization of repetitive calculations, such as the probability of infection for each individual in a mixing group. OpenMP is a useful approach because it allows the developer to annotate preexisting simulator source code with compiler directives for parallelization. The same source code still works when compiled for in single-core systems and with appropriate the annotations the basic algorithms are clearly expressed even for the parallelized versions. However, because of the overhead of multicore threading, such as starting multiple threads, dividing work, and collecting results, it is not possible to achieve a meaningful speed up at all simulation stages, especially with few infections. The evaluation results show that using an excessive number of cores could lead to less run-time performance, and that it could be more efficient in terms of throughput to limit the number of cores used and run more than one simulation job (scenario) simultaneously than to use all available cores for a single job. Nevertheless, it is possible to achieve massive improvements by dynamically switching between single and multicore mode as the simulated outbreak evolves and the computational load varies.

---

1The GNU compiler collection (GCC) provides partial support for OpenACC since version 5. In addition, there are more mature commercial compiler solutions for OpenACC from PGI, Cray, and CAPS as well as compilers developed as part of research efforts, such as OpenUH and OpenARC.
Acknowledgements

The Swedish Research Council supported this work under contracts 2008-5252 and 2009-6291.

References

Integrating the patient portal into the health management work ecosystem: user acceptance of a novel prototype

Jordan Eschler¹, Perry Lin Meas¹, Paula Lozano, MD, MPH², Jennifer B. McClure, PhD², James D. Ralston, MD, MPH², Wanda Pratt, PhD¹
¹University of Washington, Seattle, WA; ²Group Health Research Institute, Seattle, WA

Abstract

People with a chronic illness must manage a myriad of tasks to support their health. Online patient portals can provide vital information and support in managing health tasks through notification and reminder features. However, little is known about the efficacy of these features in managing health tasks via the portal. To elicit feedback about reminder and notification features in patient portals, we employed a patient-centered approach to design new features for managing health tasks within an existing portal tool. We tested three iteratively designed prototypes with 19 patients and caregivers. Our findings provide insights into users’ attitudes, behavior, and motivations in portal use. Design implications based on these insights include: (1) building on positive aspects of clinician relationships to enhance engagement with the portal; (2) using face-to-face visits to promote clinician collaboration in portal use; and (3) allowing customization of portal modules to support tasks based on user roles.

Introduction

Patients with chronic illness have a particularly burdensome health task workload, with effective management of chronic conditions potentially requiring thousands of discrete health tasks annually¹. Patients or caregivers managing a chronic illness must juggle often complex medication regimens, comply with routine screening and testing, and attend regular medical visits². Accordingly, patients or caregivers often monitor their health status daily and communicate relevant changes to their clinicians to maintain an appropriate treatment plan.

As part of managing these frequent, important tasks, reminders and notifications can be essential mechanisms for remembering and accomplishing chronic care tasks. In this paper, reminders encompass (electronic) prompts that were generated by the health care information system and pushed to a patient portal to engage patients and caregivers with upcoming recommended care tasks. Notifications are informational updates provided by the health-care information system for patients or caregivers that enhance communication around health tasks. Because patients are more likely to engage with care tasks if associated reminders and notifications are relevant and actionable³, our study works to incorporate patient preferences with reminders and notifications to enhance engagement with chronic care tasks⁴, ⁵, ⁶, ⁷.

The patient portal offers a powerful platform to customize reminders and notifications. The portal allows patients to view aspects of their electronic health record (EHR) and communicate electronically with clinicians in their health care system. These portals are information systems that could easily incorporate patient-centered approaches, particularly to take into account patient preferences, providing a means to improve illness management by promoting patients’ engagement in their care⁸, ⁹ and adherence to care plans¹⁰, ¹¹. Research has previously demonstrated the health management benefits from patient engagement with portals for specific populations (e.g., people with diabetes¹², teenagers¹³, and people with multiple sclerosis¹⁴). However, as observed in previous qualitative research on use of the patient portal, patients and caregivers experience common frustrations, such as difficulties in following up with clinicians¹⁵, failures in personal reminder systems¹⁶, ¹⁷, and gaps in attitudes between doctors, patients, and caregivers about the use of technology in health management¹⁸.

This study focuses on management of chronic illness in particular. We recruited individuals who manage diabetes for themselves (patients), and mothers who manage asthma for a child dependent (caregivers), to study the use of portal reminder and notification features. We utilize prototype testing to explore patient and caregiver interactions with new features in an existing patient portal that are intended to improve workflow, help users to remember important health tasks, and enhance patient and caregiver engagement and agency in managing chronic illness. This paper provides insights from patients and caregivers from this portal redesign to inform design implications for patient portals. We found that patient portals can best support remembering and accomplishing chronic illness management tasks by: (1) incorporating positive aspects of the clinician relationship; (2) bridging the information worlds of the clinic visit and home-based, self-care tasks; and (3) allowing users to customize patient portal functions based on role and associated tasks.
Participants
Our population of interest consisted of current users of the existing patient portal offered by a regional health care cooperative based in Washington State, U.S.A. The portal is available for patient use and for custodial parents of children 12 and under (i.e., caregivers, who manage health information for a dependent). We chose current users of the patient portal so participants would be familiar with and understand expected functionality of the system. We invited participants from two cohorts: those who manage a chronic illness for themselves (patients with diabetes mellitus), or for at least one minor child (mothers of children with asthma). For patients with diabetes (n=12, average age 63.4 years), we identified and invited individuals who had a diagnosis of diabetes mellitus. Each cohort received a paper prototype appropriate to their role: the diabetes cohort tested a patient-facing portal, and the asthma cohort tested a portal designed for caregivers of minor children. Detailed demographic information for each cohort is shown in Table 1.

Methods
We tested paper prototypes designed specifically for the tasks included in the study. Low-fidelity paper designs represented features in the portal interface, with the study facilitator acting as a computer receiving input and simulating output. We chose such low-fidelity prototypes because participants would be less likely to think of such a prototype as “finished,” and therefore would be more likely to suggest changes. In addition, paper prototypes facilitate rapid iteration and require less “polish” to deploy when testing over short periods of time. We observed participant interactions and behavior with the new features through the prototype, and audiotaped participants using a “think aloud” protocol that gave us additional data to understand how they were thinking about system interactions. The process of our prototype development is discussed in greater detail below.

Procedures
Each session was conducted individually with a participant. We used a set of four tasks to test prototype reminder and notification features that remained consistent between prototype iterations. Test sessions were audio recorded to verify participant quotes and settle disputes between researcher notes, if necessary. To maximize reliability of data collection among participants, we used a session checklist and a structured questionnaire, which was used at the end of the session to elicit additional, unstructured feedback after the task-based activities; we quote participant feedback from the questionnaire in the results. Individual sessions lasted between 30 and 45 minutes, facilitated by one or two research assistants. The three prototypes were tested with 19 individuals: seven participants tested the first prototype (P1), six participants tested the second prototype (P2), and six participants tested third prototype (P3). Most sessions were hosted in clinic conference rooms; however, some participants (n=3 mothers of children with asthma) preferred to participate in their own homes, in which case only one of the authors facilitated the session in their homes. We reached saturation in discovering design flaws after approximately six users, meaning we received similar feedback from multiple participants, which proved sufficient to iterate on the prototype. Each participant was compensated $25.

We recorded data about participant engagement with tasks in two ways. First, tasks were noted as either complete (meaning the participant followed through with the task using the prototype) or incomplete (meaning the prototype had shortcomings that prevented the participant from following through with the task). Second, we recorded other qualitative feedback from the questionnaire and think-aloud audio data to assess whether the participant accepted the task or not. Specifically, when a patient accepted the task, they made remarks that the task would be helpful or

<table>
<thead>
<tr>
<th>Table 1: Demographics of participant sample, by cohort and overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Average age (years)</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>High school</td>
</tr>
<tr>
<td>Some college/AA</td>
</tr>
<tr>
<td>4-year college</td>
</tr>
<tr>
<td>&gt;4-year college</td>
</tr>
<tr>
<td>Race</td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Native Hawaiian or Pacific Islander</td>
</tr>
<tr>
<td>More than one race reported</td>
</tr>
<tr>
<td>White</td>
</tr>
</tbody>
</table>
useful; when a patient rejected a task, they made remarks about why the task was inappropriate or inapplicable to them. We articulated these two dimensions to help determine (1) how tasks related to participants’ preferences (accepted/rejected tasks) and (2) when the prototype performance failed (complete/incomplete tasks). That is, it was possible for a participant to reject a task, but complete it, or to accept a task, but fail to complete it.

Description of tasks
In each session, the participant was presented with the home screen of the prototype and asked to talk about what s/he observed and which features s/he noticed, etc. From there, a researcher would move into facilitating four tasks, in the following order: The Lab Result Task, The Urgent Care Task, The Flu Shot Task, and The Custom Reminder Task. The tasks required an increasing amount of “work” as the user learned the system through interaction. The tasks were also designed to represent various levels of task familiarity and different information behaviors (e.g., retrieving, updating). See Table 2 for a detailed description of each of these task dimensions.

The Lab Result Task asked the user to look up a recent medical test result, and initiate a follow-up with their doctor to discuss. Patients have indicated in previous research⁴ that they did not always understand the content of test result notifications and were compelled to use secure messaging or a phone call to request further information.⁵ We refer to this patient preference as a desire to “close the loop” in communication. This task presented a common scenario with an expanded, customized feature that expedited requesting clinician commentary on a given test result, and facilitated extended “tracking” of the issue if necessary.

Table 2. Dimensions of Four Prototype Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Familiarity of task*</th>
<th>Information behavior facilitated by task</th>
<th>Change to current system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Result Task: Retrieve laboratory test results and message doctor to follow up with questions and concerns</td>
<td>High; essentially builds on existing secure messaging function</td>
<td>Open a thread of communication with the doctor via the portal to assess a specific issue</td>
<td>Reduces cognitive load in transferring information to a blank message by autocompleting text to specify issue</td>
</tr>
<tr>
<td>Urgent Care Task: Retrieve information about a prescription from an urgent care visit</td>
<td>High; introduces new view of existing, provided medication information</td>
<td>Retrieve information from the portal about medication instructions</td>
<td>Digitizes information retrieval about medication</td>
</tr>
<tr>
<td>Flu Shot Task: Dismiss a reminder about getting a flu shot by submitting up-to-date information from outside the system</td>
<td>Medium; allows participants to interact with existing flu shot reminders from portal</td>
<td>Add information to the medical record to dismiss a reminder from the portal</td>
<td>Allows patients to directly update medical record information, rather than requiring a clinician to update</td>
</tr>
<tr>
<td>Custom Reminder Task: Add custom, system-generated reminders for a planned clinic visit directed by the physician</td>
<td>Low; applies patient preferences to deploy more relevant and actionable system reminders</td>
<td>Add information to customize reminders issued by the portal</td>
<td>Empowers patients to set custom reminders for completing tasks related to personal goals</td>
</tr>
</tbody>
</table>

*Tasks are deemed more familiar if they are similar to current portfolio functionality and less familiar if they introduce new or significantly revised functionality to the portal.

The Urgent Care Task presented a task as follows: verify instructions for a medication prescribed during a recent urgent care visit. For the diabetes patients, users looked up medication instructions for themselves. For the mothers, users looked up instructions on behalf of their child with asthma. Here, we are assessing patients’ acceptance of managing comprehensive medication information online (a feature that is not available on the current patient portal), as well as exploring the type of information that would be most useful to support patient and caregiver management of medication information between clinic visits. This task adds new functionality to the portal, essentially increasing the number of ways to view medical record information about medication history.

The Flu Shot Task presented a scenario where the portal presented a reminder to get a flu shot for the patient or child, but that individual had already received a flu shot outside of the health care cooperative (and therefore the medical record had missing information). The task then requested that the patient or caregiver dismiss a flu
vaccination reminder in the system by inputting information about the flu shot that was already received. This task facilitated two new user abilities: dismissing reminders, and inputting or correcting information in the EHR through the patient portal.

**The Custom Reminder Task** asked users to add to the portal their own custom reminder that would help them to follow through on completing a task with their doctor. The task was presented as follows: your doctor told you/your child to return every six months to manage a health concern; determine when you/your child should go back to the doctor and set a reminder for yourself in the patient portal. Patients and caregivers often manage these reminders in their own calendar/notification ecosystem\(^{16}\), and with this task we tested participants’ acceptance of a feature that would assist them in integrating the portal system into their personal ecosystem of reminders.

In summary, the **Lab Test Result** and **Urgent Care** tasks introduced new ways to complete familiar patient portal tasks, while the **Flu Shot** and **Custom Reminder** tasks introduced new functionality and expanded participants’ abilities in the system. In the following section, we describe the prototype iterations.

**Iterative stages of the prototype**

We created the initial prototype designs specifically for this study based on our previous qualitative research focusing on patient and caregiver management of reminders and notifications and related patient portal use\(^{15, 16, 17, 18}\). Prototype design and iteration is described in more detail below. For all three prototypes, we increased the level of personalization on every screen by displaying the user’s name and photo on the home screen. For mothers of children with asthma, this personalization component also served as the method to toggle from one user to another. By selecting the mother’s picture, or the child’s picture, the user could navigate between profiles, and receive system feedback about which profile was “activated” depending on which name was displayed. Both patients and caregivers were very accepting of these touches of personalization, though opinions about uploading a picture—“like on Facebook,” as several of our participants noted—varied quite a bit. Mothers tended to indicate they would use the pictures (“I’d love to see my kids’ faces,” per A02), though others indicated they would prefer to use an avatar. Patients with diabetes were similarly mixed about their preference to use a picture. Because the patient/caregiver picture feature was generally well received, we retained that feature for all three prototypes. However, other features changed according to participant feedback.

Finally, although the participants interacted with paper “screens” using all three prototypes, using their fingers to complete tasks, the screens themselves were laid out digitally and were of a fairly high-fidelity layout. Feedback detailed in the below sections is from mothers of children with asthma (A# identifiers) or patients with diabetes (D# identifiers). The following sections detail the features of prototypes one through three.

**Prototype 1 (P1)**

We built the prototype’s functionality to facilitate completion of the tasks, allowing users to learn the system. The functionality of these prototypes was not very “deep,” but it gave us a chance to test a much simpler interface than presented by the current patient portal site. The primary features of P1 (see Figure 1) that differed from the current patient portal are as follows:

- Reminders and notifications were centered on the home screen and new items were highlighted in red;
- The menu taxonomy was simplified from the current patient portal menu options;
- The large buttons to navigate through the site were optimized for touch-screen use, rather than a desktop layout (most of our participants use their phones or tablets to interact with the current patient portal application); and
- The “Goals” feature was displayed, but not built out, to start a conversation with participants about what that portion of a portal might entail, and how it might benefit them.

The most common criticism of P1 was that the secure messaging option was now buried in other features. For example, when completing the Medication Task, the
participant needed to navigate to previous test results and then select “Message my care team” to use that feature. However, both cohort’s participants indicated that secure messaging should always be one click away, on every screen, because it was a “top-level” task – i.e., a primary reason for visiting the portal in the first place. Thus, access to the secure message was prioritized in the next two prototypes.

Other feedback about the prototype praised the “simple layout,” which was contrasted against the current portal’s “busy” interface. We retained a relatively simple, touch-friendly layout in subsequent prototypes, but noted that incomplete tasks with P1 stemmed from poor navigation taxonomy. For example, “Tests and care” was too vague for participants to use efficiently. Thus, we overhauled the navigation menu in P2.

Prototype 2 (P2)
We re-organized the menu options (see Figure 2) to respond to the criticisms of P1 that (1) secure messaging should be available at the top level of navigation and (2) that navigation taxonomy needed to use clearer categories. We also added redundant navigation to a side menu, which participants liked because it conveyed “consistency.” One other major change to P2 from P1 was changing the “push” notification indicators. The participants did not understand the list of notifications being front and center on the home screen, as it was in P1: participants preferred for less obtrusive notifications and reminders. We changed the design accordingly, appending relevant push notification indicators to corresponding menu options. Here, we used an iOS-like red number notification to draw the participant’s attention to that area, informed by participant input that they liked the notification styles of their phones and tablets. Participants regularly showed us their own devices during testing, and many used smartphones.

Criticisms of P2 stemmed mainly from the mothers’ use of the system: these participants indicated that they needed more information on the screen. Specifically, mothers stated that they frequently called the clinic while using the portal, particularly when they “hit a wall” in completing a task on the portal. For this reason, mothers wanted their child’s member number and the clinic phone number displayed on every screen. We made this change in P3.

Prototype 3 (P3)
The iteration from P2 to P3 was subtler than the previous redesign (see Figure 3). Most of the changes were made to accommodate mothers’ information needs and behavior, e.g., adding the health cooperative member number and clinic phone number to all screens in the patient’s profile. Another subtle change from P2 was that of using “active” language in the navigation menu to help participants understand how they would use the menu options, which had proved a shortcoming of P2. To remedy this confusion, we changed the wording of menu options to facilitate understand of the tasks supported in each of the site areas. For example, “Upcoming Care” became “Plan a Visit,” to help patients understand that they could still make appointments in the portal. “Care History” also became “View Care History,” to convey that this portion of the site served the “information archive” function that we had heard was a vital part of using the site, particularly for mothers.
Results: Task Outcomes
The following results express the outcomes by each of the tasks, which are summarized in Figure 4. When participants accepted a task, they indicated that being able to carry out the task on the portal was useful in managing health care, and they would be likely to complete the task in the manner presented in the prototype. Sometimes, though, participants did not accept tasks, saying that they would never use the portal for the task as described, or that they preferred to use whatever current solution they had in place, rather than the portal. We had numerous instances where a participant would complete the task but not accept it. In reporting on each of the tasks below, we report on both aspects of the participant feedback to maximize the analysis of the qualitative data collected during the testing sessions. All participant feedback is attributed to asthma mothers (A# identifiers) or patients with diabetes (D# identifiers).

The Lab Results Task
The user acceptance of this task was mixed, in that most of the participants saw the value of the enhanced messaging functionality, but some were reluctant to use the new feature over other communication approaches (in person or by phone). Incomplete tasks using P1 overwhelmingly stemmed from navigation issues. This led to clearer menu options in the second and third prototypes. Participants who succeeded in completing this task using P2 and P3 reviewed the test result, and then messaged the doctor using a new feature that populated a message with a subject heading and some text. A07 stated, “I like the fact that it automatically gives you an option to message your care team based on something really specific...I usually end up copying and pasting [text].” This function reduces patients’ need to remember details about the test to compose (from scratch) a message to their doctor. Incomplete task attempts in P2 and P3 resulted from the participants skipping the step to review test results and instead message the doctor their inbox in the portal. When asked about making this choice, participants stated that they mostly used the patient portal for messaging, and generally headed directly for the inbox. There, they would compose a message that would refer to the test in question, but not integrate details from the test results; instead, they would ask the doctor to interpret the result in prose in a reply. Despite incomplete attempts at the task, overall feedback from this feature was positive. Many participants said they would add some personal text to the message, to carry out the conventions of their offline relationship and to “warm up” the asynchronous communication mode.

The Urgent Care Task
Acceptance for this task was high, as many participants indicated they used the portal as an information archive (e.g., D09, who printed out archived information from the portal as needed). Patients with diabetes tended to like the medication view immediately; many of our participants managed many prescriptions, and drug interactions were a serious concern for them. These patients indicated that they would like to be able to view and edit their medications: “If it were my system,” D05 stated, she would work to “clean up” the medications list and make sure all of the information was accurate. Most existing research excludes patients as stakeholders to medication modules in health information technology, but detailed research about offering medication management tools directly to patients is lacking.

<table>
<thead>
<tr>
<th>Lab Test Results Task</th>
<th>Urgent Care Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>A majority of participants accepted this task (79% acceptance)</td>
<td>A majority of participants accepted this task (79% acceptance)</td>
</tr>
<tr>
<td>Task completion increased between prototype iterations</td>
<td>Task completion increased between prototype iterations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flu Shot Task</th>
<th>Custom Reminder Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>A majority of participants rejected this task (37% acceptance)</td>
<td>A majority of participants accepted this task (79% acceptance)</td>
</tr>
<tr>
<td>Task completion decreased between prototype iterations</td>
<td>Task completion decreased between prototype iterations</td>
</tr>
</tbody>
</table>

Figure 4: Acceptance and completion outcomes for each of the four prototype test tasks
Mothers indicated they preferred a digital After Visit Summary (which is a paper artifact already used in the health care cooperative system) to manage medications. To increase acceptance of this task, we built redundant navigation in the design iteration to P3. This satisfied both the participants with diabetes, who wanted a medication view, and the mothers managing asthma, who wanted contextual information around prescription medication in the form of the After Visit Summary. In sum, successful completion of this task was relatively high, with factors resulting in incomplete tasks resolved once the redesign for P3 incorporated the mothers’ preferences in navigation.

**The Flu Shot Task**

This task was not well accepted. Participants declined to accept the task based on their rejection of the flu shot as an important task (e.g., one participant, D07, refused one every year as a matter of course; “[the nurse] made a note of it that I didn’t want one and so they don’t do that…I just ignore [flu vaccine reminders]”). In P1, most incomplete tasks resulted from navigation errors. In general, we found, participants had a hard time figuring out where flu shots – or vaccinations in general – “fit” in the patient portal relative to other types of care.

P2’s design was much more successful, due to improved menu options, although most of the participants rejected the task itself, primarily due to rejection of the flu shot in general (as stated above) or an established habit of messaging with the care team to update information on the patient or caregiver’s behalf. In the last case, two of the six participants pointedly refused to add information directly to the medical record, and elected instead to message the doctor directly.

Participants in P3 who did not complete the task overwhelmingly opted to message with their doctor, rather than update flu shot information or dismiss the reminder. Participants stated a preference to have clinic staff update medical record information on their behalf. The primary takeaway is that participants did not necessarily make the connection between having updated medical information at hand and inputting that information into their record, and in fact saw the direct input of information as a “bypass” of the official information channel (clinic staff).

**The Custom Reminder Task**

This task was different from the first three for two reasons: (1) the functionality the system presented in completing the task was new and more unfamiliar to the participants than the first three tasks and (2) patients have existing—and sometimes complex—personal reminder systems, that they might be reluctant to integrate with what they view as an outside system. In the test of P1, most of the incomplete tasks were navigation-related. In P2, participants were more enthusiastic about using the design and chose to send themselves redundant reminders over multiple communication approaches (e-mail and text) to assist in remembering future care tasks. The results in P3 were similar to P2, with the exception that the diabetes patients testing P3 stated they would not use such a feature because they had frequent phone contact with a diabetes nurse from their clinic, and did not require extra reminders.

In the following section, we discuss themes identified from the qualitative data gathered during the testing sessions, and list design implications for patient-centered interventions in patient- and caregiver-facing health information systems.

**Discussion: Themes in Participant Feedback**

The information we elicited from prototype testing not only assisted us in iterating on the portal feature prototypes, but also provided rich qualitative data around the attitudes, beliefs, and behaviors of patients and caregivers as they integrate the use of the portal tool into their day-to-day health management routines. Lim et al. describe prototypes as “a tangible attempt to view a design’s future impact so that we can predict and evaluate certain effects before we unleash it on the world.” The authors go on to outline two purposes for prototyping with users: (1) exploring possible design decisions to manifest iterative simplifications toward a finalized design, and (2) to help designers determine what design dimensions are relevant or irrelevant. Our use of paper prototypes focused primarily on the latter purpose. Below, we elaborate on observed themes during participant interactions with the prototypes.

**The clinician relationship affected portal use**

Sometimes, tasks didn’t resonate with participants due to their strategies for maintaining (or avoiding) communication with clinicians they liked or disliked. D05 stated, “If you have a good relationship with your doctor you can work out goals that are reasonable.” For example, patients who had frequent communication with their doctors (these participants tended to describe their diabetes as “not well-controlled”) trusted their doctors to explain test results in each notification. D10 described using secure messaging: “Even if I ask maybe a dumb question [via secure message], [my doctor] answers it still. He’s very understanding.”

On the other end of the clinician-patient relationship spectrum, one mother (A04) had had a particularly poor experience with her doctors and care team. For this reason, she stated she would never attempt to message about a
test result. Instead, she preferred to use an Internet search (“I would Google something like, ‘when do I start worrying about...’”) to educate herself about the test, rather than initiate an exchange with her doctor over the portal. Previous research upholds the influence of clinician-patient relationships on the use of communication systems. Design implications for the patient portal related to this finding call for patient portal features that can amplify the positive aspects of the patient-clinician or caregiver-clinician relationship. Portal features that could potentially benefit patients who like their doctors and/or care teams potentially include: clinician pictures accompanying secure messages, to put a face to the words in secure messages; or a dedicated resource on the portal to view individual clinician or care team profiles. To accommodate those users with varying levels of rapport with their clinicians, the patient portal could incorporate user preferences to these suggested features (i.e., a caregiver with a poor relationship with clinicians could choose to hide or de-prioritize the care team profiles).

**Context of portal use differed between cohorts**
Although it is easy to assume that participants with diabetes were not technologically savvy because they were older than the mothers of children with asthma, many participants in the diabetes cohort used touch screen devices to use the existing patient portal. One participant (D10) indicated that she “liked to explore” new websites. However, some participants were particularly security conscious: “Security and whatnot...I keep no passwords on my phone whatsoever” (D08). At least one participant in the diabetes cohort used a third-party application for managing medications and remembering dosages and timing (D03).

This participant, and many of the diabetes cohort participants, proudly demonstrated use of personal devices during the course of the testing session, which we found to be helpful information, rather than a consequence of topic drift. Most participants in the diabetes cohort were eager to interact with a redesigned patient portal. We found that although the patient portal can incorporate health tasks more easily, and many patients use the technology perfectly well, it might not be the patient’s preference to do so, due to factors surrounding use. Health literacy, digital literacy, and tool use have a complicated relationship in this sense, as supported by previous research.

Mothers managing asthma on behalf of a child indicated they were busy and totally focused on the child, often prioritizing their children’s health needs over their own health tasks. Diabetes patients, however, tended to be older, and many had retired. These patients were most focused on their own health, and had numerous points of contact with the medical system (pharmacy, laboratory for regular blood tests, nurses, and often multiple physicians managing comorbid conditions accompanying diabetes). While mothers needed to plan ahead for clinic visits, diabetes patients expressed contentment with the social aspects of completing health tasks at the clinic. D12 stated, “I’m more the old school and I’ll call people up on the health team...primarily I find it works better based on my upbringing.”

Patients also felt as though they do not have a relevant ownership stake in portal information, or essentially that the doctor or other clinical staff is “in charge” of that information. Thus, as a consequence of the use context, patient portal use is subject to the power dynamics of the traditional doctor/patient dyad. It is important to recognize that this clinician/patient or clinician/caregiver collaborative stance will probably not be fostered solely through the patient portal approach, but instead requires integration of multiple approaches to encourage patient portal use to benefit both patient or caregiver, and the clinician.

The most enlightening of our findings with regard to context of use was that participants tended to trust their existing reminder systems, and expressed wariness about integrating yet another tool (e.g., the portal) into these systems. Here, we suggest, the design implication for patient portals may be offering a function that assists clinicians in encouraging patients to investigate the patient portal, to bridge the information worlds of the clinic and the home. For example, at the close of a clinic visit, a clinician could suggest that the patient use the portal to track visit follow-up tasks. At this point in the face-to-face visit, if the clinician had access to a module in the EMR to walk a patient or caregiver through portal reminder and notification features, this may prompt the user to re-engage at home.

Using such a module, the clinician could offer to set up a recurring reminder through the portal for the patient to engage with a follow-up task. By offering a visible demonstration of collaboration with a clinician (e.g., a medical assistant) during an in-person encounter, the patient or caregiver may be better able to see the relevance of portal functionality in managing care tasks, and therefore more willing to integrate the portal into their existing reminder ecosystems.
**Workflows depended on role and illness managed**

Our participants were not all simply “patients” or “caregivers.” Mothers were, of course, patients themselves and could use the portals for their own health needs. In some cases, diabetes patients were also caregivers themselves. One participant (D09) looked after her parents, in their 80’s, and did not drive. Most of her electronic calendar was used to track their appointments, and she often neglected herself to look after them: “[Taking care of them] makes me less conscientious about myself, because I’m so focused on making sure they’re taken care of. I feel like a mom.”

In contrasting workflows among users in different roles, it seems insufficient to simply modify the patient portal for caregiver use, although this is the prevailing model for offering custodial parents access to a child’s portal. Design implications of this finding might entail offering a level of customization to the patient portal. For example, users could select frequently used “modules” to highlight on the front page of the portal. Diabetes patients might prioritize modules for (1) alerts about standing orders for blood tests coming up, to keep the patient on schedule; and (2) medications management view. A caregiver of a child with asthma may prioritize (1) the asthma management plan assigned by the doctor, which assigns medication instructions to control asthma symptoms; and (2) archived information, such as summaries of emergency room visits, that the patient could easily share with the doctor to follow up with emergent asthma attacks. Allowing all users to prioritize functions of the patient portal that they use most would maximize flexibility of the patient portal.

**Limitations**

This research was conducted with patients in a single regional health-care system, with participants who had familiarity with the use of patient portal technology. We suggest future work related to patient portal redesign or implementation should incorporate contextual and social aspects of in situ tool use\(^\text{35}\), which our lab-based paper prototype testing did not include.

**Conclusion**

Health management tasks are particularly important and time-consuming for patients and caregivers who are managing a chronic illness. To support these tasks, we designed and tested three prototypes for a new patient portal that incorporates patient-centered reminders and notifications. The information we elicited from prototype testing with our participants helped us to identify participant attitudes, beliefs, and behavior that affect design dimensions for patient-centered reminders and notifications to manage health tasks. First, patient portal design should incorporate features to amplify positive aspects of the clinician relationship, which will benefit those users who feel a personal connection at the clinic and can carry over to the feelings of responsibility and accomplishment related to small, everyday tasks. Second, clinicians should be provided with an easy way to gauge interest in portal use, and encourage patients and caregivers to use the portal to bridge the information worlds of the clinic and the home, where many complicated care tasks occur. Finally, customization of the patient portal modules can provide a flexible platform that supports the roles of both patients and caregivers, offering better solutions for both. These design implications contribute evidence-based insight as to the needs of patient and caregiver stakeholders in patient portal tools. This study gives rich user feedback to inform design opportunities in patient portal systems that can effectively support the sometimes overwhelming tasks of patients and caregivers who manage chronic illness every day.

**References**

Modeling the Temporal Evolution of Postoperative Complications

Shara I. Feld, PhD¹, Alexander G. Cobian, MS²,³, Sarah E. Tevis, MD¹, Gregory D. Kennedy, MD PhD¹, Mark W. Craven, PhD³,²
¹Department of Surgery, ²Department of Computer Sciences, ³Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

Abstract

Post-operative complications have a significant impact on patient morbidity and mortality; these impacts are exacerbated when patients experience multiple complications. However, the task of modeling the temporal sequencing of complications has not been previously addressed. We present an approach based on Markov chain models for characterizing the temporal evolution of post-operative complications represented in the American College of Surgeons National Surgery Quality Improvement Program database. Our work demonstrates that the models have significant predictive value. In particular, an inhomogenous Markov chain model effectively predicts the development of serious complications (coma longer than a day, cardiac arrest, myocardial infarction, septic shock, renal failure, pneumonia) and interventional complications (unplanned re-intubation, longer than 2 days on a ventilator and bleeding transfusion).

Introduction

Patients who suffer from post-operative complications have higher rates of post-operative morbidity and mortality, resulting in longer hospital stays, higher rates of readmission and higher cost of care [1-8]. While many studies have evaluated pre-operative predictors of complication development [1, 9-16], recent work has shown that we can extend our analysis to post-operative predictors. Tevis et al. [17] found that there are relationships in the occurrence of post-operative complications. Wakeam et al. [18] found that there are associations between complication timing and mortality. However, prior research has not investigated how sequences of multiple complications develop given post-operative events. Tevis et al. [17] employed a Bayesian network approach that revealed dependencies among multiple complications, but did not account for the ordering or timing of post-operative complications. The analysis of Wakeam et al. [18] used post-operative complications to predict mortality, but did not extend to analyzing cascades of complications.

The goal of this study was to assess whether we could accurately model the temporal evolution of postoperative complications. Specifically, we evaluated how well the likelihood and timing of complication development given prior post-operative complications could be represented using Markov chain models, a well established class of models for characterizing temporal and sequential data [19].

Methods

Complication data

The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database includes 30-day postoperative outcomes for patients who underwent major inpatient and outpatient surgical procedures [20]. ACS NSQIP data are collected by a trained surgical clinical reviewer at each site, and subsequently audited for reliability by the NSQIP program. This study includes inpatient and outpatient surgical cases from over 435 institutions during the years 2005 to 2013.

The complications analyzed were the 21 reported ACS NSQIP complications occurring within 30 days after the operation and post-operative mortality (Table 1). Each recorded complication was annotated with the number of days after the operation that the complication was first diagnosed. If a complication was diagnosed multiple times post-operatively, only the first date of diagnosis was recorded. Complications included infectious types (superficial surgical site infection (SSI), deep SSI, organ space SSI, wound disruption, urinary tract infection, sepsis and septic shock), physiologic complications (peripheral nerve injury, pneumonia, deep vein thrombosis or thrombophlebitis, pulmonary embolism, renal insufficiency, renal failure, stroke or cardiovascular incident, myocardial infarction,
cardiac arrest, or coma longer than 24 hours), and interventional complications (unplanned intubation, cumulative ventilator-assisted respiration greater than 48 hours, bleeding transfusion up to 72 hours post-operatively, and graft failure requiring intervention).

Given the de-identified nature of the ACS NSQIP data, work with this dataset has been deemed exempt by the University of Wisconsin Health Sciences IRB.

<table>
<thead>
<tr>
<th>Table 1. Frequency of individual complications, among all patients diagnosed with at least one post-operative complication.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complication</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Infection</td>
</tr>
<tr>
<td>Superficial SSI</td>
</tr>
<tr>
<td>Deep SSI</td>
</tr>
<tr>
<td>Organ SSI</td>
</tr>
<tr>
<td>Wound Disruption</td>
</tr>
<tr>
<td>UTI</td>
</tr>
<tr>
<td>Sepsis</td>
</tr>
<tr>
<td>Septic shock</td>
</tr>
<tr>
<td>Physiologic</td>
</tr>
<tr>
<td>Nerve injury</td>
</tr>
<tr>
<td>Pneumonia</td>
</tr>
<tr>
<td>DVT</td>
</tr>
<tr>
<td>Pulmonary Embolism</td>
</tr>
<tr>
<td>Renal insufficiency</td>
</tr>
<tr>
<td>Acute renal failure</td>
</tr>
<tr>
<td>Stroke/CVA</td>
</tr>
<tr>
<td>MI</td>
</tr>
<tr>
<td>Cardiac Arrest</td>
</tr>
<tr>
<td>Coma&gt;24 hrs</td>
</tr>
<tr>
<td>Intervention</td>
</tr>
<tr>
<td>Unplanned Intubation</td>
</tr>
<tr>
<td>On ventilator&gt;48 hrs</td>
</tr>
<tr>
<td>Bleeding transfusion</td>
</tr>
<tr>
<td>Graft failure</td>
</tr>
<tr>
<td>Death</td>
</tr>
</tbody>
</table>

Markov models

Markov models allow us to represent the temporal evolution of complications as patients progress through series of health states [19]. We constructed discrete-time Markov chain models to characterize the development of subsequent complications given knowledge of the complications a patient has already experienced. To develop our Markov models, we defined states that represent the patients’ health condition and estimated the transition probabilities of changing from each state to another. The outcomes that we predicted were the occurrence of each of the 21 post-operative complications and mortality. The time scales of the data and model are one day steps, from the day of surgery to 30 days post-operation.
Figure 1. (a) The state representation used by our Markov model. The top part of the panel shows the new complications that are recorded on each of three days in a given patient’s record. The bottom part of the panel shows the vector of binary variables representing the patient’s state on each day. (b) The graphical model representation of our Markov model. Each node corresponds to a random variable representing the status of a given complication on a specific day. Each dashed box delineates the set of variables that represent the state for a given day. The edges impinging on each node represent the potential dependencies of the corresponding variable on the complications from the previous day.

We define a patient’s health state at a given point in time by the set of complications they have experienced prior to this point in time. With this definition, we represent a state using a vector of binary variables, one per complication, where a given variable is set to 1 if the corresponding complication has already occurred, and is set to 0 otherwise. An example using this representation is shown in Figure 1(a). The NSQIP data does not specify the durations of complications or multiple occurrences of a given complication in a patient’s record. Therefore, we designed our state representation to indicate, for a given day, all of the complications that have occurred in the patient up to that day.

Since there are 21 complications (excluding death), there are $2^{21}$ possible states. We use a factored state and transition representation for our models, as shown in Figure 1(b), and thus our models can also be considered dynamic Bayesian networks. Let $C_t$ represent the vector of random variables characterizing the complication state at time $t$, $C_t[i]$ indicate the variable for the $i^{th}$ complication at time $t$, and $c_t$ and $c_t[i]$ represent the assignment of values to the vector and to the $i^{th}$ variable, respectively.\footnote{We use bold symbols to indicate vectors, uppercase letters to indicate random variables, and lowercase letters to indicate values of random variables.}

We consider two different representations for state transitions. In both of them, the probability of a complication occurring at time $t$ is conditionally independent of other complications occurring at time $t$, given the previous state, $c_{t-1}$:

$$\Pr(c_t | c_{t-1}) = \prod_i \Pr(c_t[i] | c_{t-1}).$$

A patient’s history can be viewed as a sequence of state transitions. Since the NSQIP data records events at the time granularity of days, the state transitions in our model occur on a daily basis. In our first representation for state transitions, we use a set of logistic regression models as follows. The probability of the $i^{th}$ complication variable being 1 at time $t$ is given by:
\[
\Pr(C_t[i] = 1|e_{t-1}) = \begin{cases} 
1 & \text{if } C_{t-1}[i] = 1 \\
 \frac{p^{(i)}(e_{t-1})}{1 + e^{-(w^{(i)}e_{t-1} + b^{(i)})}} & \text{otherwise}
\end{cases}
\]

where

\[
p^{(i)}(e_{t-1}) = \frac{1}{1 + e^{-(w^{(i)}e_{t-1} + b^{(i)})}}.
\]

That is, if the variable \(C_{t-1}[i]\) is set to 1, indicating that the \(i^{th}\) complication had occurred by the previous time step, then it must remain in this state. Otherwise a logistic regression model is used to predict the probability of the complication occurring at time \(t\). The logistic regression model uses the state from the previous day, \(e_{t-1}\), for its predictor variables. Note that the model is specific for the \(i^{th}\) complication and hence we use the superscript \((i)\) for the parameters of the model, \(w^{(i)}\) and \(b^{(i)}\). The parameters in the logistic regression models were estimated using the maximum likelihood estimation algorithm implemented in the Matlab 2015a library \textit{glmfit} [21].

A second representation for state transitions we consider is based on the Noisy-Or function using the method of Onisko et al. [22]:

\[
\Pr(C_t[i] = 1|e_{t-1}) = \begin{cases} 
1 & \text{if } C_{t-1}[i] = 1 \\
1 - \prod_{j\neq i, C_{t-1}[j]=1} \left(1 - \frac{1 - p^{(j)}(e_{t-1})}{1 - p^{(i)}(e_{t-1})}\right) & \text{otherwise}
\end{cases}
\]

where the product in the second case ranges over the complications that have been set to 1 on the previous day, and \(p^{(i,0)}\) and \(p^{(i,1)}\) are defined as follows:

\[
p^{(i,0)} = \Pr(C_t[i] = 1|C_{t-1}[0] = 0, \ldots C_{t-1}[20] = 0) \quad \text{and} \quad p^{(i,1)} = \Pr(C_t[i] = 1|C_{t-1}[0] = 0, \ldots C_{t-1}[j] = 1, \ldots C_{t-1}[20] = 0).
\]

In other words, \(p^{(i,0)}\) represents the probability of the \(i^{th}\) complication occurring on day \(t\), given that no other complications have yet occurred, and \(p^{(i,1)}\) represents the probability of the \(i^{th}\) complication occurring on day \(t\) given that only the \(j^{th}\) complication has previously occurred.

In both the logistic regression and Noisy-Or linking functions, the time scales of the data and model are one-day steps, from the day of surgery to 30 days post-operatively. The status of complications that a patient experienced through time is represented by linking the transition functions for the likelihood of complication development over the 30-day post-operative period [19].

For both the logistic regression and Noisy-Or approaches, we consider homogenous and inhomogenous variants of the models. In the homogenous case, the data from all days is pooled and then used to infer model parameters that are invariant throughout the 30-day post-operative time period. For the inhomogenous models, we partition the data into time-based subsets and then learn model parameters that are specific to given post-operative time periods. The partitioning is done such that each period contains approximately 20% of the complications that occurred across the population. These time periods were the day of surgery, 1-3 days after surgery, 3-8 days after surgery, 8-15 days after surgery and 15-30 days after surgery.

**Results**

**Model evaluation methodology**

To evaluate the extent to which our models have captured the temporal evolution and dependencies among complications, we assessed how well they are able to predict a patient’s health state one day in advance given the post-operative complications the patient has experienced up to that point. We did this using a 10-fold cross validation methodology. The data were stratified by patient (i.e., a given patient’s health history was either entirely in the training set or entirely in the test set) and randomly partitioned. The predictions for each test set were made using models learned from the corresponding training set. The test set predictions were then pooled for a cumulative evaluation using ROC curves.
Model predictive value

The ROC curves in Figure 2 display predictive accuracy for the homogenous and the inhomogenous models that use logistic regression to represent transition probabilities. Each curve represents the true positive rate and false positive rate of the model predicting all cases (day 0 predicting day 1, day 1 predicting day 2, up to day 29 predicting day 30) for a specific complication.

The inhomogenous models had significant predictive value for almost all complications. The models were best at predicting 30-day mortality, serious complications (coma longer than a day, cardiac arrest, MI, septic shock, renal failure, pneumonia) and interventional complications (unplanned re-intubation, longer than 2 days on a ventilator and bleeding transfusion). The complications in this set were all predicted with an area under the curve greater than 0.85. The other complications are more difficult to predict because they are likely to occur in isolation, and are not as dependent on prior complications. For example, surgical site infections are dependent on wound closure and post-operative wound management, which are not characterized in the NSQIP dataset, rather than other complications.

![Figure 2. ROC curves for the logistic regression-based models showing the true positive rate (TPR) against the false positive rate (FPR). Solid lines show predictive accuracy for the inhomogenous models, heavy dashed lines represent the homogenous models, and the light dashed lines represent the expected curves for random guessing (i.e., no predictive value).](image)

Comparison of homogenous and inhomogenous models

As indicated in Figure 2, the inhomogenous models had superior predictive accuracy to the homogenous models. For all complications, there was improved predictive accuracy in using the homogenous models, and for some, the improvement was dramatic. This result was true for both the models that used logistic regression and those that used Noisy-Or to represent transition probabilities. This result indicates that the transition probabilities and
dependencies vary according to the time elapsed since surgery. Not only does sequencing matter when predicting whether a complication is likely to occur, but how far out the patient is from surgery.

The improvement in predictive accuracy that comes from using inhomogenous models indicates that modeling sequences of post-operative complications requires incorporating multiple factors that influence complication development. Some complications depend on specific timing of occurrence – for example, bleeding transfusions are by design only recorded within the first 72 hours. Other complications may occur in isolation or not in a predictive fashion, such as superficial SSI. In other cases, the inhomogenous model better characterizes transition probabilities that vary throughout the post-operative period.

Comparison of logistic regression and Noisy-or models

The predictive accuracy of the logistic regression and Noisy-Or models was similar in both the inhomogenous and homogenous cases. Figure 3 shows a comparison of these two different approaches for representing transition probabilities. Each point in the figure represents a specific complication, and the coordinates of the point are given by the area under the ROC curves for the Noisy-Or and logistic-regression models. The tight clustering of the points along the diagonal indicates that the two approaches have similar predictive accuracies for every complication.

![Figure 3](image)

**Figure 3.** Scatterplots comparing area under the ROC curves for the logistic regression and Noisy-Or models. Each point represents the area under the ROC curves for a specific complication. Shown are the (a) homogenous comparison and (b) inhomogenous comparison.

How relationships between complications are highlighted

The models can be used to highlight the strength of relationships among the complications that may be anecdotaly understood but not formally measured. For example, the inhomogenous logistic regression model, as depicted in Figure 4, highlights that stroke/cardiovascular event, cardiac arrest, and prolonged period on the ventilator are most predictive for developing a coma. The inhomogenous model is able to parse out the temporal variation in the extent to which predictor complications contribute to the development of the dependent complications. The inhomogenous model further highlights how complications such as surgical site infections (SSI), that may have appeared non-contributory in the homogenous model, can serve as indicators for more serious future complications at different stages of the post-operative course.
Figure 4. Parameters for the logistic regression models that predict the occurrence of the complication, coma > 24 hours. Shown are the parameters for both the homogenous model (•) and the inhomogenous model (bars). In the inhomogenous model, there are separate parameters for each time period: days 0-1 post operation, days 1-3, days 3-8, days 8-15 and days 15-30.

Conclusion

Our empirical results tell us that Markov models can effectively represent risk for later complications depending on the complications a patient has already experienced. We found that complications that are strongly dependent on a given timing (such as bleeding transfusion), dependent on prior complications (such as mortality), or both are modeled well. Complications that occur in isolation and in a wide range of times (such as superficial surgical site infection) are modeled poorly. While Tevis et al. [17] demonstrated that there are relationships among which post-operative complications occur, the present work demonstrates that those relationships are often sequential and thus have prognostic power for assessing patient outcome. The models further indicate the extent to which complications are prognostic for the subsequent occurrence of dependent complications.

Developing the model required considering timing after surgery in addition to which predictive complications have occurred. Prior work indicates that Markov models work well in applications in which the timing of events is important, which is beneficial in medical decision making [19]. We found that an inhomogeneous model, in which dependencies are conditioned upon when the events occurred over the 30-day post-operative period, outperformed a homogenous model in terms of predictive accuracy. This supports prior work noting that complication occurrence after surgery varies depending on time after surgery [24].
The standardization and size of the ACS NSQIP database also strengthens our study. The data is collected in a prospective manner by trained surgical clinical reviewers, complications are strictly defined, and the national database has a large patient population which allows us to assess the temporal dependencies in the <5% of surgical patients included who developed multiple complications.

This study has limitations inherent to the dataset and to our Markov models. We were limited by the prevalence of complications. The dataset does not indicate the durations of complications, and it only includes the first date on which a complication is diagnosed even if the complication is experienced multiple times. Furthermore, the complication diagnosis data is recorded at a temporal resolution of one day, excluding modeling dependencies between same day complications.

The limitations of our models include their first-order nature and their inability to explicitly represent dependencies between complications co-occurring on the same day. We do not view the first limitation as significant given that our state representation encodes the history of complications experienced up to the given day. Our logistic regression versus Noisy-Or comparison suggests that the inability to represent dependencies between same-day complications does not significantly hinder predictive accuracy either. Although the logistic regression models cannot explicitly represent such dependencies, they are able to compensate for their effects somewhat. Our experiments demonstrated that the logistic regression models did not offer any predictive advantages over the Noisy-Or models which assume that there are no such dependencies. Our models could be extended in a number of ways in order to capture more information about the temporal evolution of complications. Some of the more promising extensions would be to use explicit duration modeling, and to incorporate other NSQIP variables (demographics, surgical procedures, comorbidities, etc.) in the models.

Acknowledgments

This work was supported by NIH/NIDDK grant T35 DK062709, NIH/NLM grant T15 LM07359, NIH/NCATS grant UL1 TR000427, BD2K program NIH/NIAID grant U54 AI117924, and the University of Wisconsin Department of Surgery.

References

Towards Comprehensive Clinical Abbreviation Disambiguation Using Machine-Labeled Training Data

Gregory P. Finley, PhD¹, ², Serguei V.S. Pakhomov, PhD¹, ³, Reed McEwan, MS, MSSE¹, Genevieve B. Melton, MD, PhD¹, ²

¹Institute for Health Informatics, ²Department of Surgery, and ³College of Pharmacy
University of Minnesota, Minneapolis, MN

Abstract

Abbreviation disambiguation in clinical texts is a problem handled well by fully supervised machine learning methods. Acquiring training data, however, is expensive and would be impractical for large numbers of abbreviations in specialized corpora. An alternative is a semi-supervised approach, in which training data are automatically generated by substituting long forms in natural text with their corresponding abbreviations. Most prior implementations of this method either focus on very few abbreviations or do not test on real-world data. We present a realistic use case by testing several semi-supervised classification algorithms on a large hand-annotated medical record of occurrences of 74 ambiguous abbreviations. Despite notable differences between training and test corpora, classifiers achieve up to 90% accuracy. Our tests demonstrate that semi-supervised abbreviation disambiguation is a viable and extensible option for medical NLP systems.

Introduction

The frequent use of abbreviations in clinical texts is a major challenge for natural language processing (NLP) systems. Medical abbreviations are especially challenging because they tend to be highly ambiguous: a 2001 survey reports that nearly a third of shorter abbreviations catalogued in the Unified Medical Language System (UMLS) are ambiguous;¹ a later study reveals that even the UMLS sense inventory provides spotty coverage of all possible abbreviation senses.² The problem is even worse in clinical notes created for patient care, as opposed to other biomedical texts mostly derived from peer-reviewed literature, due to issues ranging from higher word ambiguity³ to mistakes in spelling and dictation.⁴ Even detecting which words are abbreviations is not a trivial task, and modern clinical NLP systems have much room for improvement in this area.⁵

Resolving ambiguities in clinical texts is a major concern for improving medical information retrieval outcomes.⁶,⁷ As such, an active area of research in medical NLP is in the normalization of abbreviations.⁸ For ambiguous abbreviations, this is generally treated as a special case of the word sense disambiguation (WSD) problem: determining the sense of a single string that may have multiple distinct semantic interpretations.⁹,¹⁰ Supervised machine learning approaches to abbreviation disambiguation based on WSD techniques have been generally successful.¹¹, ¹² However, most prior studies in this domain have evaluated only a handful of abbreviations; the generalizability of these results to more realistic use cases, with hundreds or thousands of abbreviations, is limited.

Fully supervised methods are also subject to the limitation that they require training data that is slow and expensive to obtain. Furthermore, to keep up with the rapid and ongoing proliferation of clinical abbreviations, such data would require constant maintenance. To obviate the need for labeled data, some researchers have proposed unsupervised¹³ or knowledge-based¹⁴ strategies to the disambiguation problem. Still, published research in this area generally focuses on fewer than 20 abbreviations.

Another possibility is a semi-supervised (or “distantly supervised”) approach, which requires some attention to data collection but not nearly the commitment necessary for fully hand-annotated data. Normalizing abbreviations differs from other WSD problems in that the senses (i.e., long forms) have distinct string realizations. This property can be exploited to generate examples of virtual occurrences of an abbreviation by targeting its various long forms in a search of unlabeled text. This general approach has been developed for biomedical texts¹⁵ and clinical notes,¹⁰,¹⁶ with reported accuracy approaching 90%. Stevenson et al. apply a similar corpus generation process to Medline abstracts that have abbreviations co-occurring locally with their long forms.¹¹ (Note that these convenient co-occurrences are rare for clinical notes.) They employ a battery of features to achieve 99% accuracy for 20 abbreviations.

The main objective of the present study was to conduct a large-scale test of the semi-supervised method just described. Ours differs from prior research in two important ways. First, training and test data are drawn from
different corpora. This point is key, as training and testing on homogeneous data is not a realistic evaluation of an NLP application for many cases—versatile clinical NLP systems should be useful for a wide variety of texts, not just those similar to data on which its models were trained. Other studies on semi-supervised biomedical acronym normalization which train and test on different corpora are by Pakhomov et al., who disambiguate 8 acronyms with up to 67.8% accuracy,16 and Xu et al., who disambiguate 13 abbreviations from typed hospital admission notes with up to 87.5% accuracy.12

Second, we test a wider range of abbreviations than has been considered previously: 74 ambiguous abbreviations frequent in clinical texts but not specific to any particular field of medicine. The only other published studies that test a comparable number all focus on the same abbreviations that we do (or a subset of them).17–19 These studies, however, rely on fully supervised cross-validation on a single corpus rather than separately generated training and test sets. By applying a semi-supervised approach to a wide scope of abbreviations in a more realistic test case, we present the most extensible method to date for disambiguating abbreviations in clinical texts. (The test set we use actually comprises 74 initialisms, commonly referred to in the literature as “acronyms,” which are a sub-type of abbreviation that involves the concatenation of initial letters of words in a phrase. We continue to use the term “abbreviation” because our methods should be equally applicable to acronyms and to other abbreviations.)

The semi-supervised method is appealing because it should maintain the high accuracy and efficiency of fully supervised approaches without requiring expensive human tagging of corpora. It does depend, however, upon an assumption that the distributions of an abbreviation and of its long forms are similar. This assumption seems intuitively well supported, but at the same time it is easy to imagine reasons why distributions might differ—abbreviated forms may tend to occur more often in notes more reliant on abbreviations, or a writer may consciously avoid abbreviations in contexts where they would be too opaque or unusual.

That said, it is not evident a priori that these weaknesses in the assumption lead to unsatisfactory results. The present study is an empirical evaluation of the semi-supervised method, simulating a real-world use case: training a normalization system on machine-annotated data and testing it on natural occurrences of a wide range of abbreviations.

Methods

Corpora and sense inventory

Two corpora were used in this study. The first was taken from the Clinical Abbreviation Sense Inventory.20 This publicly available data set (“CASI”) lists 440 common clinical abbreviations along with their long forms, mappings to medical concepts, and other information. Most of these abbreviations were found to have a dominant sense accounting for 95% or more of 500 randomly sampled occurrences; for those 74 that do not, CASI provides anonymized plaintext data and manually annotated senses for 500 samples of each (37,000 total). These samples, minus the 223 marked as having ‘unsure sense’, constitute Corpus A.

Corpus B was built by querying a large clinical data repository in the Fairview Health Services system (about 90 million notes) for long forms of these abbreviations using Elasticsearch. (All notes in the health record were searched, so this corpus contains results from many different types of documents across numerous specialties. We used definitions of notes as based off of the HL7-LOINC document ontology and as catalogued in our data repository.) In most cases, unmodified long forms themselves could serve as queries. A few, however, required slight adjustments. For example, searching for the long form ‘computed tomographic angiography’ returned very few results; the slightly modified (case-insensitive) query ‘ct angiography’ returned many more, while still being unambiguous. In another case, the expansion ‘gutta’ (for ‘GT’) was rare in the corpus (and ambiguous with a physician’s name), while the semantically equivalent ‘gtts’ was more common. Overall, we kept any modification of the searches to a minimum; queries were only altered if an informal examination of the search results showed them to be scant, overly repetitive, or obviously incorrect.

827,647 total notes were returned by queries across 207 different senses, with a maximum of 5,000 notes per sense. Samples were taken from the body text of these notes by identifying the long form in them and matching (greedily) any appearances of that form both preceded and followed by 40 to 100 characters. This step effectively excluded any examples with too little context, such as those from very short notes. Up to 8,000 samples were generated for each sense (allowing for multiple tokens of a sense per sample). In all, a total of 857,724 annotated virtual abbreviations were generated.
The sense inventory used to collect documents for Corpus B was built from CASI. 357 total senses are represented in the data set, although 150 are exceptionally rare (occurring as fewer than 5 of its abbreviation’s 500 samples) and were excluded; in all, these account for 0.6% of the total data. (Note that examples of these rare senses were retained in the test set and thus contributed to the errors of a model not trained on them.) The procedure for generating both corpora is visualized in Figure 1. Histograms showing the number of senses per abbreviation and the number of training samples per sense in Corpus B are given in Figure 2a and 2b, respectively.

**Word vectors**

Feature vectors for each instance of an abbreviation were calculated as simple co-occurrence counts within a window of text, a “bag-of-words” representation. Tokens were lemmatized, case was ignored, and numerals were collapsed into broad categories: single-digit integers, multi-digit integers, and decimals. By these criteria, Corpus A has 13,877 unique words, and Corpus B 72,514. (Note that Corpus A has been anonymized with generic strings whereas B has not, so names, addresses, etc. contribute to the latter’s word count.)
Although word order was not taken into account explicitly, counts were weighted to increase the contribution of words in closer context. Weights were computed by a sigmoid function $W(d)$ that decreases with the context token’s (always positive) distance $d$ in words from the token of interest:

$$W(d) = \frac{1}{1 + e^{a(d-r)}}$$

where $r$ is the number of words away at which weight drops to half and $a$ controls the steepness of the curve. Various parameters for the rate and point of falloff were tested; good values were found to be around $r = 9$ with a shallow falloff of $a = 0.3$. Words weighted at less than 0.25 (greater than 12 words away for these values) were not counted at all. This weighting maintains a minor effect of word order and was found to slightly improve classification accuracy overall. Beyond about 9 or 10 words, a larger window does not generally improve accuracy for this kind of task.\textsuperscript{17}

Researchers have suggested various other features to enhance supervised classification—salient N-grams,\textsuperscript{11} knowledge sources,\textsuperscript{14} neural word embeddings.\textsuperscript{19} The central question here, however, is not how best to tailor features to maximize performance; rather, we are concerned with whether the machine-annotated data can effectively serve as a training corpus given minimal assumptions and adjustments. Bag-of-words features are simple, general, and quite reliable, and they should give a fair indication of whether it is worth the effort to further tune models and features for these data.

**Machine learning algorithms**

Several well-known classification algorithms were implemented, all using the features described above:

- **Naïve Bayes (NB):** A single Gaussian was fitted for each feature across all examples of a given class (sense). Class priors were not considered, as these would favor more frequent classes in the training data, and the relative frequencies for long forms may not be representative of sense frequencies for abbreviations—recall the earlier discussion that some senses are more likely to be abbreviated than others. Hypotheses at test time were chosen by maximizing likelihood, with consideration only of the non-zero entries in the test vector. Given the data’s high dimensionality and natural-language origins, many probabilities at test time come out at or near zero, so natural log probabilities were kept to a minimum of $-100$ to prevent a single unusual or unseen word from dooming a candidate altogether. A separate model was trained for each of the 74 acronyms.

- **Multinomial logistic regression (LR):** A classifier was trained for every abbreviation. The intercept term typical of logistic regression was not used, for the same reason class priors were not used for NB. Models were trained by 100 iterations of gradient descent, which usually converged in 3 to 10 iterations. Regularization generally did not improve results and was not used in the reported tests.

- **Support vector machine (SVM):** This classifier was identical to LR but with a standard hinge loss function ($\varepsilon = 1$) rather than log loss.

- **Cosine similarity (COS):** Test items were classified by maximizing the cosine of the angle between the test context vector and a normalized vector representing the centroid of all training items of each sense. Two other adjustments were made to these centroid sense vectors before normalization. First, all values were compressed via a square root, which effectively favors less common words and therefore diverse lexical contexts (note that this would have little effect for other classifiers, which do not average training examples). Second, words in training and testing context vectors were weighted by their inverse document frequency (IDF). The cosine similarity metric is not a discriminative or maximum-likelihood strategy, so it is essential to reduce the impact of words with low predictive power.

**Other features: hyperdimensional indexing**

Another method that has received some recent attention has been the use of high-dimensional vectors to represent words. Rather than unitary representation as a single dimension in a context vector, word vectors themselves are “hyperdimensional,” having on the order of a few thousand dimensions. (Note that word vectors are actually of lower dimensionality than the “one-hot” vectors used for a traditional word vector space, but that these representations depend on encoding several of these dimensions, whereas the one-hot vectors are predictable from the vocabulary.) Context vectors are represented by “bundling” word vectors, typically through a sum or similar operation. Word vectors are not orthogonal as unitary word vectors in a traditional vector space are, but are all
nearly orthogonal due to their high dimensionality. Context vectors are equidimensional with word vectors, so they are actually shorter (but less sparse) than the context vectors considered for the other classifiers. Two hyperdimensional approaches were tested:

- Random indexing (RI): Each unique word is represented as a ternary vector of 1,800 elements, with a random four values set to 1, another random four to –1, and the rest to 0. These are the same parameters described by Kanerva et al. Other hyperparameter values were explored but did not significantly improve classification. Context vectors are created by summing word vectors within the context window (9 tokens on either side) subject to the same IDF weighting used for the vector space model. Senses were chosen by the same similarity criterion as the COS classifier described above. In the clinical domain, RI has been applied effectively towards capturing distributional similarity between long forms and their abbreviations.

- Binary spatter code (BSC): Each unique word is represented by a 10,000-dimensional binary vector, with all values randomly assigned. During training, context vectors are calculated not for the senses themselves, but for all words within 9 tokens of an abbreviation, by summing together vectors that are the “product” (bitwise XOR) of an abbreviation and its labeled sense. Word frequency is accounted for in these sums by entropy weighting, and the final vector chosen by voting on the weighted counts of ones and zeros in each dimension. For testing, contexts are “divided” (XOR again) by the abbreviation vector to recover a context vector correlated with the most likely sense. The implementation hews very closely to the description given by Berster et al.; see also Kanerva for a theoretical discussion and Moon et al. for its suitability to abbreviation normalization.

Test conditions

Recall that this study considers two corpora derived from clinical texts—Corpus A, smaller and hand annotated; and Corpus B, larger and automatically annotated. We performed tests with each as the training set and each as the test set. Whenever a single corpus was used for both training and test, 10-fold cross-validation was performed.

The original aim of the study was to evaluate an abbreviation normalizer that is not dependent on fully supervised data. Corpus B can be created cheaply, while A contains actual occurrences of abbreviations; thus, training on B and testing on A is the intended use case and gives the best indication of performance on real-world clinical data.

To mitigate problems related to overdeveloping to this use case, we halved Corpus A into development and validation sets. Examples were split by stratified random sampling: for every sense of every abbreviation, half of the examples selected at random went to each set, with the validation set receiving the remainder for odd counts. All refinements of the models were performed according to performance on the development set (18,278 samples total), and all statistics reported in this paper are from the validation set (18,499 samples total). For conditions other than training on B and testing on A, this split was ignored and Corpus A was used in its entirety.

We have made code for the experiments in this study available online (https://github.com/gpfinley/towards_comprehensive). Corpus A is freely available; however, Corpus B contains protected health information and cannot be publicly shared.

Results & discussion

Classification accuracy

The results for classifiers tested on Corpus A are reported in Table 1. Scores reported in the left column are results of 10-fold cross-validation, applied in a classically supervised manner to Corpus A. The right column of the table represents the intended use case of the semi-supervised method.

Baseline accuracy for tests on Corpus A is defined as the majority sense—the score obtained by simply guessing the most common sense for each abbreviation. Recall that no classifiers tested in this study explicitly take sense frequency into account.

The results of cross-validation on Corpus A confirm that traditionally supervised methods are highly accurate. Hyperdimensional approaches were slightly less accurate but still reasonable. For the semi-supervised case, however, all classifiers suffered somewhat. The drop in performance, as compared to the fully supervised cross-validation results, was modest for classifiers operating on word count vectors, with COS losing only 6.1% accuracy.
Table 1. Classification accuracy for classifiers tested on the hand-annotated Corpus A. Scores in the left column are averages of 10-fold cross-validation. Baseline is the average majority sense score for all abbreviations.

<table>
<thead>
<tr>
<th></th>
<th>CV on A</th>
<th>Train on B, test on A</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>.949</td>
<td>.854</td>
</tr>
<tr>
<td>LR</td>
<td>.966</td>
<td>.880</td>
</tr>
<tr>
<td>SVM</td>
<td>.964</td>
<td>.887</td>
</tr>
<tr>
<td>COS</td>
<td>.961</td>
<td>.900</td>
</tr>
<tr>
<td>RI</td>
<td>.945</td>
<td>.817</td>
</tr>
<tr>
<td>BSC</td>
<td>.935</td>
<td>.761</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td>.735</td>
</tr>
<tr>
<td>chance</td>
<td></td>
<td>.398</td>
</tr>
</tbody>
</table>

Recall the assumption that abbreviations and their long forms share the same distribution in text. Given that this assumption is certainly not guaranteed, it can be said that LR, SVM, and especially COS perform fairly well to any variation introduced by training on long forms and testing on short forms. Corpus B was built with very light supervision, so other sources of noise probably also contribute to the drop in accuracy.

The hyperdimensional classifiers, on the other hand, showed a sharper drop in performance, up to 17.4%. Lexical differences between the two corpora may be responsible: high-dimensional word vectors are not perfectly orthogonal, causing some overlap in context vectors that would be difficult to account for if some words are present in one corpus but not another. Alternatively, it may be that hyperdimensional context vectors are especially sensitive to other differences between corpora that are unimportant for abbreviation disambiguation.

Whatever the explanation, it is clear that the RI and BSC algorithms suffer more from using machine-annotated training data than those classifiers based on the vector space model. Hyperdimensional approaches certainly have some advantages for representing lexical semantics in NLP, such as compactness and flexibility, but they appear ill suited to the task here (especially because maintaining context vectors for abbreviation senses requires less storage than for an entire vocabulary). This result is especially noteworthy given the excellent results reported for BSC in fully supervised abbreviation disambiguation by Moon et al.\textsuperscript{18}

The same classifiers were also evaluated using Corpus B as a test set, with results shown in Table 2. Note that this is a highly artificial case, as the notes used to derive Corpus B contain the long forms rather than ambiguous abbreviations. Note also that a majority-sense baseline is less meaningful for Corpus B (it would be 54.6%), which targeted the long forms directly rather than sampling actual occurrences of the abbreviation, and thus is not thought to be representative of actual sense frequency.

Table 2. Classification accuracy for classifiers tested on the hand-annotated Corpus B. Scores in the right column are averages of 10-fold cross-validation.

<table>
<thead>
<tr>
<th></th>
<th>Train on A, test on B</th>
<th>CV on B</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>.717</td>
<td>.959</td>
</tr>
<tr>
<td>LR</td>
<td>.742</td>
<td>.990</td>
</tr>
<tr>
<td>SVM</td>
<td>.748</td>
<td>.988</td>
</tr>
<tr>
<td>COS</td>
<td>.804</td>
<td>.982</td>
</tr>
<tr>
<td>RI</td>
<td>.762</td>
<td>.934</td>
</tr>
<tr>
<td>BSC</td>
<td>.610</td>
<td>.948</td>
</tr>
<tr>
<td>chance</td>
<td></td>
<td>.349</td>
</tr>
</tbody>
</table>
Fully supervised cross-validation was even better for Corpus B than Corpus A, a fact which may be due to the sheer size of Corpus B, the frequent use of boilerplate language in the source notes, or even altogether redundant notes, which were not screened out when searching the database.

In the reverse case, training on Corpus A and testing on B, no classifier performed exceptionally well (the COS classifier achieves the highest accuracy with 80.4%). Such a condition—normalizing virtual abbreviations in texts originally lacking them—is unlikely to arise in practical use. However, it may speak to fundamental differences between the corpora. While it is true that Corpus B is much larger, it is unlikely that this discrepancy alone leads to poor performance, given the finding by Moon et al.\textsuperscript{17} that a training set a fraction of the size of Corpus A can still give decent results for abbreviation normalization. A more likely explanation is that Corpus A is simply less general and diversified than B: data in Corpus A are primarily verbally dictated and transcribed, whereas Corpus B contains notes created through dictation as well as a variety of combinations of preformed templates, macros, typing, and voice recognition software. Variation inherent to Corpus A is present in B, but not vice-versa. (Note also that the impact of domain transfer may be overstated in these results because we have not controlled for the differences between hand- and machine-labeled annotations; i.e., instances of abbreviations in Corpus A are not found using semi-supervised methods.)

In terms of practical application, the performance discrepancy between the two experiments emphasizes the importance of using well-diversified data to train robust WSD systems. The semi-supervised approach, in conjunction with access to a large database of clinical notes, handles this with ease.

**Error analysis**

We also investigated several of the abbreviations that were responsible for the accuracy lost between the fully supervised and semi-supervised conditions (i.e., the two columns of Table 1). Our focus is specifically on the COS classifier, as it had the best performance. All abbreviations with an accuracy differential between the two tasks of at least ten percentage points are shown in Table 3.

**Table 3.** Differences in the accuracy of the COS classifier for the semi- and fully supervised cases. Only abbreviations with a performance differential of at least 10 percentage points are shown. Sense counts exclude very rare senses.

<table>
<thead>
<tr>
<th>abbrev.</th>
<th>senses</th>
<th>B/A acc.</th>
<th>A/A acc.</th>
<th>diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOM</td>
<td>2</td>
<td>.425</td>
<td>.998</td>
<td>.573</td>
</tr>
<tr>
<td>DT</td>
<td>3</td>
<td>.673</td>
<td>.956</td>
<td>.283</td>
</tr>
<tr>
<td>T1</td>
<td>3</td>
<td>.672</td>
<td>.944</td>
<td>.272</td>
</tr>
<tr>
<td>PAC</td>
<td>4</td>
<td>.726</td>
<td>.962</td>
<td>.236</td>
</tr>
<tr>
<td>PCP</td>
<td>4</td>
<td>.713</td>
<td>.944</td>
<td>.231</td>
</tr>
<tr>
<td>AB</td>
<td>3</td>
<td>.760</td>
<td>.958</td>
<td>.198</td>
</tr>
<tr>
<td>NAD</td>
<td>2</td>
<td>.745</td>
<td>.942</td>
<td>.197</td>
</tr>
<tr>
<td>RT</td>
<td>3</td>
<td>.773</td>
<td>.962</td>
<td>.189</td>
</tr>
<tr>
<td>PA</td>
<td>4</td>
<td>.777</td>
<td>.948</td>
<td>.171</td>
</tr>
<tr>
<td>DC</td>
<td>4</td>
<td>.755</td>
<td>.904</td>
<td>.146</td>
</tr>
<tr>
<td>IT</td>
<td>6</td>
<td>.780</td>
<td>.905</td>
<td>.125</td>
</tr>
<tr>
<td>ER</td>
<td>3</td>
<td>.856</td>
<td>.974</td>
<td>.118</td>
</tr>
<tr>
<td>GT</td>
<td>3</td>
<td>.825</td>
<td>.940</td>
<td>.115</td>
</tr>
<tr>
<td>CD4</td>
<td>3</td>
<td>.865</td>
<td>.972</td>
<td>.107</td>
</tr>
<tr>
<td>RA</td>
<td>3</td>
<td>.855</td>
<td>.962</td>
<td>.107</td>
</tr>
<tr>
<td>BMP</td>
<td>3</td>
<td>.761</td>
<td>.866</td>
<td>.105</td>
</tr>
</tbody>
</table>
For most of these abbreviations, it is not immediately evident why the semi-supervised classifier performs poorly. For a handful, though, some educated guesses can be made:

- **MOM**: Nearly all of the errors on this abbreviation were misidentifying the sense ‘multiples of median’ as ‘milk of magnesia’. The former long form was difficult to target in the notes corpus, as it is usually abbreviated, so there were few training examples for what happened to be the most common sense of ‘MOM’.

- **T1**: ‘thoracic (level) 1’ was frequently misidentified as ‘T1 (MRI)’, which was difficult to target as a long form. We ultimately designed the query for the latter to match ‘T1’ in close context with ‘MRI’, and it appears to have erroneously matched several uses that refer to the first thoracic vertebra in an MRI context.

- **PAC**: ‘physician assistant certification’ was often missed. This sense almost always follows a name, and a key difference between Corpus A and Corpus B is that the former has had all names collapsed as part of de-identification.

- **IT**: The pronoun ‘it’ was frequently identified as an abbreviation. The contexts for ‘it’ are exceptionally diverse, so it may have been difficult to motivate ‘it’ over other hypotheses. These errors could easily be sidestepped if tokens were tagged for part of speech, as most medical NLP systems do, by assuming that pronouns are not abbreviations.

It is fair to say that certain senses of these abbreviations are represented in Corpus B differently from in A. Some differences arise from terms that rarely appear as long forms (e.g., ‘multiples of median’), whereas others may reflect difficulties inherent to the semi-supervised approach (‘T1’, ‘PAC’).

Some errors, particularly in this latter category, could be substantially reduced with more attention paid to the searches used to retrieve data. The intent for this project was specifically not to do so; the goal was to determine how well minimally managed data collection would work to train machine learning models. The results obtained here for the semi-supervised case should be considered a conservative estimate of performance, and more systematic attention to queries will likely increase accuracy further.

**Extending to other abbreviations**

Our test set encompasses hundreds of senses from 74 unique abbreviations. If a system were required to only account for these 74, then a fully supervised approach might be the best recommendation. Of course, this is but a fraction of known medical abbreviations. Though our coverage is still incomplete, the semi-supervised methods that we validate here are extensible to other abbreviations and their known senses. More comprehensive sense inventories could be built either from existing medical knowledge sources, such as the SPECIALIST Lexicon, or by using unsupervised machine learning methods, such as the clustering approach proposed by Xu et al.

Evaluating the accuracy beyond these 74 is difficult without acquiring more hand-annotated test data. Nevertheless, there is no reason to believe that other abbreviations should be less suitable for the semi-supervised approach. Our tests are also rather conservative, in both the minimal treatment of database queries and the use of very general machine learning methods.

Generally speaking, better performance on primary NLP tasks such as WSD should improve outcomes in secondary tasks. While an in-depth discussion of WSD applications in clinical NLP is beyond the scope of the present paper, we would like to note briefly that the ability to identify context-appropriate senses of ambiguous medical terms, abbreviations in particular, is fundamental to several higher-level NLP and information retrieval tasks, including accurate concept mapping to standardized vocabularies and conceptual indexing of electronic health records. Fully supervised systems are feasible only for a small and static set of ambiguous terms; however, clinical notes are “living” documents whose authors often do not follow conventions and abbreviate medical terms ad libitum for convenience and speed of documentation. Unsupervised or semi-supervised systems of the kind described in this paper offer the potential to keep up with these new ambiguous patterns dynamically introduced into clinical documentation. It may even be of interest to have models configured to text in specific clinical domains, especially if there are idiosyncrasies in documentation style between providers or specialties.

**Comparison to other published results**

A subset of the data from Corpus A was also used for training and testing by Moon et al. and Moon et al.; the former tested Naïve Bayes and SVM classifiers, the latter a hyperdimensional binary spatter approach. Figures reported previously in the current study are on all 74 abbreviations of the published data set. When testing on only
the 50 abbreviations considered in the other two papers, our accuracy results are extremely similar: marginally lower for BSC (our 92.9% versus their 93.5%) and higher for NB and SVM (our 94.2% and 96.2% versus their 93.7% and 93.9%).

Note that the above comparison is only for fully supervised methods. Comparisons to other tests of semi-supervised methods are difficult to make because most other studies draw training and test items from the same corpus, which simplifies the problem considerably. The closest comparison is with Xu et al.,12 who investigate 13 abbreviations that have at least one disease sense. They draw training data from dictated discharge summaries and test data from typed hospital admission notes. Their methods for generating the corpora are nearly identical to ours: training samples are built by targeting long forms directly in a text search, while test samples are manually annotated. As such, some performance comparisons can be reasonably made. Our peak performance at 90.0%, using only bag-of-words features (with some positional weighting), slightly exceeds theirs at 87.5%, which uses a combination of bag-of-words features, positional features, section headings, and sense frequency information. Xu et al.’s reliance on sense frequency in particular may introduce bias because frequencies are calculated off of the test corpus itself, albeit with unsupervised methods. (We use sense frequency only initially to determine if any senses are vanishingly rare, and not as a feature for classification.) Xu et al. report accuracy of 79.2% when not using sense frequency.

**Limitations**

There are a few limitations to this study that should be considered. First, we do rely somewhat on sense frequency in Corpus A to exclude rare senses (less than 1% of a given abbreviation’s occurrences) from consideration. Including these senses would have required the acquisition of much more training data and would probably have had some negative impact on accuracy. Some degree of medical expertise would be helpful to determine which senses are very rare.

Second, several of the database queries that we perform do require some attention and modification. The amount of labor is significantly less than that required for fully supervised labeling of examples, and it does not scale with the size of the data set. Nevertheless, queries optimized for one database and search protocol might not transfer well to others.

Finally, implementing the semi-supervised method with clinical notes as training data requires access to an appropriately large database, which many members of the larger research community do not have. Sharing models trained on clinical note data, such as the ones in this study, is difficult due to confidentiality concerns and would require removal of potentially identifying information.

**Future work**

Though our results show some success with semi-supervised methods, a more thorough investigation should be conducted to quantify the actual distributional similarity between abbreviations and their long forms and to understand how the assumption of similarity affects the results. Our own preliminary investigation has found that this assumption generally holds but is much more valid for some abbreviations than for others. Distributional similarity may depend on a number of factors—such as text domain, writing style, and accidental homonymy with other terms—and may interact with NLP task accuracy in interesting ways.

**Conclusion**

In this study we demonstrated the suitability of a semi-supervised approach for disambiguating abbreviations in clinical text on a large scale. Our tests covered more data and more distinct abbreviations than have been tested previously by this method, and our test conditions are a better representation of real-world problems. Generating machine-annotated training data requires little human supervision and can be easily extended to other abbreviations. Applying this method has the potential to significantly improve information retrieval outcomes for clinical text.

**Acknowledgments**

The National Institutes of Health through the National Library of Medicine (R01LM011364 and R01GM102282), Clinical and Translational Science Award (8UL1TR000114-02), UMN Academic Health Center Faculty Development Award, and Fairview Health Services supported this work. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.
References


569
Leveraging Lexical Matching and Ontological Alignment to Map SNOMED CT Surgical Procedures to ICD-10-PCS

Kin Wah Fung1, MD, MS, MA; Julia Xu1, MD, PhD; Filip Ameye, MD, FRCS(Eng), FACS; Arturo Romero Gutiérrez2, MD; Arabella D’Havé, MSc, BSc3

1National Library of Medicine, Bethesda, MD; 2Ministry of Health, Social Services and Equality, Spain; 3Federal Public Service of Health, Food Chain Safety and Environment, Belgium

kwfung@nlm.nih.gov|xujc@mail.nlm.nih.gov|filip_ameye@telenet.be|aromerog@msssi.es|arabella.dhave@gezondheid.belgie.be

Abstract
In 2015, ICD-10-PCS replaced ICD-9-CM for coding medical procedures in the U.S. We explored two methods to automatically map SNOMED CT surgical procedures to ICD-10-PCS. First, we used MetaMap to lexically map ICD-10-PCS index terms to SNOMED CT. Second, we made use of the axial structure of ICD-10-PCS and aligned them to defining attributes in SNOMED CT. Lexical mapping produced 45% of correct maps and 44% of broader maps. Ontological mappings were 40% correct and 5% broader. Both correct and broader maps will be useful in assisting mappers to create the map. When the two mapping methods agreed, the accuracy increased to 93%. Reviewing the MetaMap generated body part mappings and using additional information in the SNOMED CT names and definitions can lead to better results for the ontological map.

Introduction
After being used for over 30 years for reimbursement and other purposes in the U.S., the ICD-9-CM coding system was finally replaced by ICD-10-CM in October 2015. While procedures were included as part of ICD-9-CM (volume 3), they are not part of ICD-10-CM. To replace the ICD-9-CM procedure codes, a brand-new coding system called ‘The International Classification of Diseases 10th Revision Procedure Coding System’ (ICD-10-PCS) was created by the U.S. Centers for Medicare and Medicaid Services (CMS) through a contract with 3M Health Information Systems1, 2. ICD-10-PCS is now used in the U.S. for the reporting of inpatient procedures. Other countries that have used ICD-9-CM are also transitioning to ICD-10-CM and ICD-10-PCS. Belgium made the transition in early 2015, and similar changes will happen in Spain and Portugal in the near future. In the realm of electronic health records (EHR), SNOMED CT has been designated as the clinical terminology standard. The Meaningful Use of EHR incentive of CMS specifies SNOMED CT as the terminology for the encoding of problem lists and procedures3. Many of the member countries of the International Health Terminology Standards Development Organisation (IHTSDO) are using, or have plans to use, SNOMED CT to document medical procedures. To facilitate the interoperability and integration of clinical and administrative data, it is evident that a map between SNOMED CT procedure codes and ICD-10-PCS will be necessary.

A project group was formed under IHTSDO in early 2015 with the goal to create a map between SNOMED CT and ICD-10-PCS.4 The group includes representatives from Belgium, Spain, Portugal and U.S., together with medical terminology and informatics vendors. The map is expected to provide the following benefits:

- Improve ICD-10-PCS coding efficiency, consistency and accuracy
- Promote re-use of clinical data for administrative, epidemiologic and statistical purposes
- Facilitate integration of clinical and administrative data to support data analytics

Another potential benefit of the map is related to CPT (Current Procedure Terminology), another procedure terminology maintained by the American Medical Association (AMA). CPT is used in the U.S. for the billing of medical procedures and physician services. There is ongoing negotiation between IHTSDO and AMA to build a map between CPT and SNOMED CT. If both CPT and ICD-10-PCS are linked to SNOMED CT, SNOMED CT can become a lingua franca to facilitate interoperability between the two procedure coding systems.
Among the topics discussed by the project group are automated ways to map between the two terminologies. It is understood that the mappings generated by algorithmic methods will still need human validation, but reviewing candidate maps that are of reasonably high accuracy is easier and faster than creating a map from scratch. Among the methods considered, the group identified lexical matching of the ICD-10-PCS index terms and ontological alignment as among the most promising approaches of automatic mapping. This paper describes the two mapping methods and results of their evaluation.

**Lexical Mapping**

Various methods of automatic mapping have been studied. Generally, these are either lexical or semantic methods. Lexical methods rely on terms matching, while semantic methods depend on relationships. The axial structure of ICD-10-PCS poses special challenges and opportunities to automatic mapping.

Traditional lexical mapping depends on finding matches between names of codes from two terminologies. String normalization can improve matching rate by reducing variations due to punctuations, upper/lower cases and lexical variants such as conjugation. However, direct lexical mapping between SNOMED CT and ICD-10-PCS names does not work well because of two reasons. First, ICD-10-PCS names are composite terms created by combining the values in the various axes, which are often different from terms used in clinical discourse. For example, the ICD-10-PCS name for laparoscopic appendectomy is *Resection (or Excision) of appendix, percutaneous endoscopic approach*, which will not match the common clinical term *Laparoscopic appendectomy* used in SNOMED CT. Second, there is mismatch in granularity between SNOMED CT and ICD-10-PCS. A full ICD-10-PCS code often includes information about approach (e.g., open approach, via natural or artificial opening), device (e.g., synthetic substitute, metallic joint prosthesis) and intent (e.g., diagnostic). In SNOMED CT, such additional information is only included in a minority of surgical procedures.

On the other hand, the ICD-10-PCS index is more promising for lexical matching because it contains many ‘clinician-friendly’ terms, such as:

![Image of ICD-10-PCS codes]

**Ontological mapping**

ICD-10-PCS is built on a multi-axial structure. The seven characters in the code correspond to the seven axes, each describing a particular aspect of the procedure.

This structure makes ICD-10-PCS expandable because new procedures can be easily incorporated as unique codes. In SNOMED CT, concepts are logically defined by attributes, which is in some ways similar to an axial structure. SNOMED CT descends from predecessors that are based on an axial structure. The use of attribute-value pairs in definitions liberates SNOMED CT from the limitations of earlier axis-based definitions, but the compositional nature
of the terminology remains.\textsuperscript{12,13} By aligning the ICD-10-PCS axes and SNOMED CT attributes, it is possible to identify equivalence between the two terminologies (table 1). For example, laparoscopic appendectomy is defined in SNOMED CT by the attribute-value pairs: Method = Excision, Procedure site - direct = Appendix structure and Using access device = Laparoscope. By aligning the three attributes Method, Procedure site - direct and Using access device to Root operation (axis-3), Body part (axis-4) and Approach (axis-5) respectively in ICD-10-PCS, and matching the values for these attributes/axes, one can find the equivalent ICD-10-PCS codes 0DTJ4ZZ or 0DBJ4ZZ (figure 1).

<table>
<thead>
<tr>
<th>SNOMED CT attribute</th>
<th>ICD-10-PCS axis</th>
<th>Axis no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure site</td>
<td>Body system</td>
<td>2</td>
</tr>
<tr>
<td>Method</td>
<td>Operation</td>
<td>3</td>
</tr>
<tr>
<td>Procedure site - Direct</td>
<td>Body part</td>
<td>4</td>
</tr>
<tr>
<td>Access</td>
<td>Approach</td>
<td>5</td>
</tr>
<tr>
<td>Surgical Approach</td>
<td>Approach</td>
<td>5</td>
</tr>
<tr>
<td>Using access device</td>
<td>Approach</td>
<td>5</td>
</tr>
<tr>
<td>Direct device</td>
<td>Device</td>
<td>6</td>
</tr>
<tr>
<td>Indirect device</td>
<td>Device</td>
<td>6</td>
</tr>
<tr>
<td>Procedure device</td>
<td>Device</td>
<td>6</td>
</tr>
<tr>
<td>Using device</td>
<td>Device</td>
<td>6</td>
</tr>
<tr>
<td>Procedure site - Indirect</td>
<td>Qualifier (Site)</td>
<td>7</td>
</tr>
<tr>
<td>Has intent</td>
<td>Qualifier (Intent)</td>
<td>7</td>
</tr>
<tr>
<td>Using substance</td>
<td>Qualifier (Substance)</td>
<td>7</td>
</tr>
<tr>
<td>Direct substance</td>
<td>Qualifier (Substance)</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1. Matching SNOMED CT procedure attributes to ICD-10-PCS axes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Section: 0 Medical &amp; Surgical</td>
<td>Operation: T Resection</td>
<td>Body part: J Appendix</td>
<td>Approach: 4 Percutaneous endoscopic</td>
<td>0D[TB]J4ZZ</td>
</tr>
<tr>
<td>Body system: D Gastrointestinal</td>
<td></td>
<td>Device: Z No device</td>
<td>Qualifier: Z No qualifier</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. An example of ontological mapping, based on matching attributes and values at source and target systems. A source SNOMED CT procedure concept (1) has some attributes and values defining the concept; SNOMED CT attribute Method can be aligned to PCS target axis Operation (2); Procedure site - direct can be aligned with Body part(3); Using access device can be aligned to Approach (4)
Methods

Creating the lexical and ontological maps

From the 2016 version of the ICD-10-PCS Index xml file, we extracted main index terms and the associated codes listed with the main entries and sub-entries. We mapped the index terms to SNOMED CT concepts using NLM’s MetaMap program (2015 version).\textsuperscript{14, 15} Since the ICD-10-PCS index also included terms that were not procedures (e.g., body parts and devices), we restricted the SNOMED CT concepts to surgical procedures, defined as descendants of \textit{Surgical procedure} (387713003). We considered all ICD-10-PCS codes included under a main index term as candidate map targets for the SNOMED CT concept mapped to that index term.

For the ontological map, we first collected all the body parts values used in axis-4 of ICD-10-PCS. To this list, we added the inclusion terms, which were finer-grained terms included by the body part values (e.g., auditory ossicle includes incus, malleus and stapes), as defined by the ICD-10-PCS Definitions xml file. We mapped all body part terms to SNOMED CT using MetaMap, restricting the output to SNOMED CT body structure concepts. In this study, we focused on the ICD-10-PCS codes with excision or resection as the root operation. For all surgical procedures in SNOMED CT, we identified those with \textit{Method} = \textit{Excision}, and matched the \textit{Procedure site – direct} value to body part terms in ICD-10-PCS. If there was more than one relationship group in the SNOMED CT definition, the body part match had to come from the same group as the excision method. We considered all ICD-10-PCS codes with root operation of resection or excision, and a matching body part value, to be candidate map targets for that SNOMED CT procedure.

Evaluation of the maps

We selected a random sample of 100 concepts from each map, among which 50 concepts were common to both maps, to be reviewed by two of the authors (JX and FA). Each individual mapping was rated independently by each reviewer as:

\begin{itemize}
  \item a. Correct, no better codes available
  \item b. Too specific, should use more general codes
  \item c. Too broad, should use more specific codes
  \item d. Related somewhat, not broader/narrower, better codes available
  \item e. Target codes unrelated to source concept
\end{itemize}

If the mapping was correct, the reviewer would also determine if all correct map targets had been identified. For each incorrect mapping, the reviewer would assign a reason for failure. Results from the reviewers were compared, and the differences were discussed until a consensus was obtained. If a consensus could not be reached, a third reviewer (KWF) would cast the deciding vote.

Results

Lexical mapping

A total of 2,557 unique main entries were extracted from the ICD-10-PCS index. MetaMap mapped the index entries to 569 unique SNOMED CT surgical procedures, which were in turn mapped to 5,679 unique ICD-10-PCS codes. The majority of the SNOMED CT concepts (95\%) were mapped to four or less ICD-10-PCS codes, mostly at the 3 or 4 character level. A random subset of 100 SNOMED CT concepts was selected for review, excluding concepts with maps to over 20 ICD-10-PCS codes. In the review subset, 100 SNOMED CT concepts mapped to a total of 141 unique ICD-10-PCS codes (54 4-character and 87 3-character codes). Among them, 33 concepts mapped to only one ICD-10-PCS code, 52 concepts to two codes, with the maximum of nine codes (one concept). There were altogether 197 mappings (unique pairs of SNOMED CT and ICD-10-PCS codes) for review. Table 2 summarizes the review results for lexical mapping.
Table 2. Results of lexical mapping

Overall, 45% of the maps were correct. Among the incorrect maps, the following reasons of failure were identified:

1. **Index not sufficiently granular**

The index pointed to a 3- or 4-character code, but the correct map was more specific. For example, the index for choledochectomy led to 0F[B/T], but the correct map should be 0F[B/T]9.

2. **MetaMap error**

MetaMap mapped to the wrong SNOMED CT concept. For example, Osteoplasty was mapped to Bone fusion.

3. **ICD-10-PCS index entry was questionable**

In some cases, we think that the ICD-10-PCS index was not correct. For example, Colpocentesis pointed to drainage of the vagina (0U9G) while the correct map should be drainage of the cul-de-sac (0U9F). Another example is Gastroenterostomy, which pointed to either bypass or drainage. We think that drainage is not the correct root operation in this case since the procedure aims at altering the route of passage of the contents of the stomach to the small bowel, rather than letting out fluids from the stomach.

Looking at the results from the perspective of SNOMED CT concepts, 44% of the concepts had at least one correct map, and 96% had at least one correct or broader map. For those concepts with correct mappings, there were no missing map targets that the map failed to identify.
**Ontological mapping**

There were altogether 19,658 surgical procedures in SNOMED CT, and each procedure was defined by at least one *Method* attribute, whose value was restricted to the *Surgical action (129284003)* sub-hierarchy. *Excision* was the most common method, followed by *Surgical action* and *Repair* (table 3).

<table>
<thead>
<tr>
<th>Method attribute values</th>
<th>Number of concepts (% of all surgical procedures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excision - action (qualifier value)</td>
<td>3788 (19%)</td>
</tr>
<tr>
<td>Surgical action (qualifier value)</td>
<td>2356 (12%)</td>
</tr>
<tr>
<td>Repair - action (qualifier value)</td>
<td>2253 (11%)</td>
</tr>
<tr>
<td>Incision - action (qualifier value)</td>
<td>1467 (7%)</td>
</tr>
<tr>
<td>Closure - action (qualifier value)</td>
<td>1118 (6%)</td>
</tr>
<tr>
<td>Reconstruction - action (qualifier value)</td>
<td>726 (4%)</td>
</tr>
<tr>
<td>Surgical removal - action (qualifier value)</td>
<td>690 (4%)</td>
</tr>
<tr>
<td>Fixation - action (qualifier value)</td>
<td>680 (3%)</td>
</tr>
<tr>
<td>Destruction - action (qualifier value)</td>
<td>672 (3%)</td>
</tr>
<tr>
<td>Grafting - action (qualifier value)</td>
<td>483 (2%)</td>
</tr>
<tr>
<td>Other values</td>
<td>5425 (28%)</td>
</tr>
<tr>
<td>Total</td>
<td>19658 (100%)</td>
</tr>
</tbody>
</table>

Table 3. Distribution of methods in the definitions of SNOMED CT surgical procedures

Ontological mapping found maps for 1,746 SNOMED CT surgical procedures, mapping to 689 unique ICD-10-PCS codes (all 4-character). 95% of the concepts were each mapped to 10 or less ICD-10-PCS codes. The review subset consisted of 100 randomly selected concepts, excluding those mapping to over 20 ICD-10-PCS codes. Altogether, the 100 concepts mapped to 237 unique ICD-10-PCS codes. Most of the concepts (82 concepts) mapped to one to four ICD-10-PCS codes. There were a total of 362 mappings for review. The results are summarized in table 4.

<table>
<thead>
<tr>
<th>Map category</th>
<th>Mappings (%)</th>
<th>Reason of failure</th>
<th>Resection vs excision</th>
<th>Operation mismatch</th>
<th>ICD-10-PCS inconsistent</th>
<th>Not classifiable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>146 (40%)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Too specific</td>
<td>83 (23%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Too broad</td>
<td>19 (5%)</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Related</td>
<td>61 (17%)</td>
<td>15</td>
<td>25</td>
<td>17</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Unrelated</td>
<td>53 (15%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>362 (100%)</td>
<td>126</td>
<td>25</td>
<td>17</td>
<td>2</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 4. Results of ontological mapping

Overall, 40% of the maps were correct. Among the incorrect maps, the following reasons of failure were revealed:

1. **Body part mismatch**

The map was incorrect due to the body part term in ICD-10-PCS being mapped to an incorrect body structure in SNOMED CT concept. In most cases, the SNOMED CT body structure was too broad, which resulted in the mapping of a broad SNOMED CT procedure to an overly-specific ICD-10-PCS code. For example, MetaMap mapped *Face artery, Thyroid artery* and *Hand artery* in ICD-10-PCS to *Arterial structure* in SNOMED CT. As a result, *Excision of artery* in SNOMED CT was mapped to *Excision of face artery, Excision of thyroid artery* and *Excision of hand artery* in ICD-10-PCS.
2. **Missing approach information**

In this study, we did not make use of information about surgical approach for mapping. Some SNOMED CT concepts containing approach information were therefore mapped to ICD-10-PCS codes that were too broad. For example, *Laparoscopic adrenalectomy* in SNOMED CT was mapped to excision/resection of left/right adrenal gland (e.g., 0GB2 – Excision of left adrenal gland), while the correct map target should include the approach (e.g., 0GB24 – Excision, left adrenal gland, percutaneous endoscopic approach). Other examples included *Transurethral prostatectomy* and *Thoracoscopic excision of neoplasm of pericardium*.

3. **Excision vs. resection**

While SNOMED CT considered excision and resection as synonymous, they were considered different root operations in ICD-10-PCS. In ICD-10-PCS, excision was defined as “the cutting out or off, without replacement, a portion of a body part”; while resection was “the cutting out or off, without replacement, all of a body part”. Generally, the correct mapping of a SNOMED CT excisional procedure should include both the excision and resection options in ICD-10-PCS. However, in cases where the SNOMED CT concept referred specifically to a partial or complete removal, a more specific ICD-10-PCS code should be used. For example, *Excision of focal lesion of larynx* should only map to *Excision of larynx (0CBS)* and not *Resection of larynx (0CTS)*. On the other hand, *Total thymectomy* should map to *Resection of thymus (07TM)* but not *Excision of thymus (07BM)*.

4. **Root operation mismatch**

In SNOMED CT, excision is used broadly to define procedures that involve removal or destruction of a body part, regardless of the method used. In ICD-10-PCS, the codes may be different depending on the method and what is being removed. For example, *Chemothalamectomy* was a destruction operation in ICD-10-PCS. *Vasectomy using silicon plug* would be coded as occlusion. *Removal of sequestra from bone* was an extirpation procedure.

5. **ICD-10-PCS inconsistent**

There was one case in which the reviewers found conflicting information in the ICD-10-PCS index. While the entry *Cricoidectomy* pointed to the excision of larynx, the entry for Cricoid cartilage suggested to use trachea as the body part.

6. **Unclassifiable**

There were three SNOMED CT concepts that we considered not classifiable in ICD-10-PCS. *Excision of accessory or ectopic lung tissue* was not classifiable because the location of the ectopic lung tissue was not known (could be in the thorax, abdomen or subcutaneous). *Excision of benign lesion of trunk* was not classifiable because trunk was not a defined body region in ICD-10-PCS. *Laparoscopy with excision of lesion* was too broad to be meaningfully classified.

Looking at the results from the perspective of SNOMED CT concepts, 68% of SNOMED CT concepts had at least one correct map, and 78% had at least one correct or broader map. Among the SNOMED CT concepts with correct maps, the map missed some target codes in four concepts because the SNOMED CT concept contained more than one procedure. One example is *Anterior resection of rectum with colostomy* in which the target code for colostomy was missed.
Comparing the two maps

Half of the reviewed concepts were common to both maps by design, so that we could compare the two mapping methods. For the 50 concepts in common, there were 101 lexical and 182 ontological mappings. We analyzed the mappings by whether they had: a. exact same target in the other map b. matching map target in the other map at the 3-character level c. no matching map targets (table 5).

<table>
<thead>
<tr>
<th></th>
<th>Number of mappings</th>
<th>Correct map %</th>
<th>Broader map %</th>
<th>Correct or broader map %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical Map (LM)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same target as OM</td>
<td>29</td>
<td>93%</td>
<td>7%</td>
<td>100%</td>
</tr>
<tr>
<td>Match OM target at 3-character level</td>
<td>53</td>
<td>19%</td>
<td>79%</td>
<td>98%</td>
</tr>
<tr>
<td>No match</td>
<td>19</td>
<td>58%</td>
<td>37%</td>
<td>95%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>101</td>
<td>50%</td>
<td>49%</td>
<td>98%</td>
</tr>
<tr>
<td><strong>Ontological Map (OM)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same target as LM</td>
<td>29</td>
<td>93%</td>
<td>7%</td>
<td>100%</td>
</tr>
<tr>
<td>Match LM target at 3-character level</td>
<td>116</td>
<td>50%</td>
<td>0%</td>
<td>50%</td>
</tr>
<tr>
<td>No match</td>
<td>37</td>
<td>14%</td>
<td>0%</td>
<td>14%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>182</td>
<td>50%</td>
<td>1%</td>
<td>51%</td>
</tr>
</tbody>
</table>

Table 5. Accuracy of maps for concepts common to both mapping methods (OM: ontological map, LM: lexical map)

Overall, the two reviewers agreed in 50% of the mappings after the first round of individual assessment. Most differences were resolved after discussion, and only 10% of mappings required the third reviewer’s vote to decide.

Discussion

Creating a map between two terminologies is labor intensive and time consuming. Mapping is not an exact science. Due to the differences in scope, granularity and organizing principles between the source and target terminologies, the mapping experts sometimes have to rely on their judgment, which can result in considerable variability. Common methods to ensure consistency include dual independent mapping (each map created by two mappers independently) and sequential review (review of some or all maps by a second mapper). Automated methods to identify candidate maps can help by suggesting map targets to mappers, which can save time and improve consistency. The candidate maps can also act as an independent source to corroborate the manually-created maps. Our previous study shows that algorithmically created maps can lead to considerable saving in mapping time.16

In our study, 45% of the maps generated by lexical mapping of the ICD-10-PCS index are correct. While this level of precision is not particularly impressive on its own, an additional 44% of the maps are broader maps at the 3- or 4-character level. In the use case of assisting manual mapping, the broader maps are also useful because they lead to the correct table, from which it is relatively simple to pick the correct target code. So we can say that 89% of the maps are useful in this context. Overall, 96% of SNOMED CT concepts which can be mapped in this way will have a useful map. However, the main drawback of this method is the limited coverage, as the index of ICD-10-PCS is not very extensive. Only 569 SNOMED CT surgical procedures can be mapped by this method.

On the other hand, ontological mapping can potentially cover many more SNOMED CT concepts. We are able to find maps for 46% (1,746 out of 3,788) of the SNOMED CT excision procedures. The precision of the ontological map (40%) is slightly lower than the lexical map, and the proportion of broader maps is also lower (5%). Overall, the ontological map is able to find at least one correct or broader map for 78% of SNOMED CT concepts.
Analyzing the reasons of failure enables us to discover ways of improving the performance of the maps. The main reason of failure in the lexical map is insufficient granularity of the codes listed in the index, most of them only at the table (3-character) level. There is no way to get around this at present, unless the ICD-10-PCS developers decide to add new entries or use more granular codes in the index in future. For the ontological map, there are some possibilities for improvement. Body part mismatch makes up the majority of failures. Our method depends solely on the output of MetaMap without manual review. Since the total number of body part terms in ICD-10-PCS is relatively small (less than 800), it is feasible to manually review and correct the maps found by MetaMap, which could improve the overall performance. Another possible improvement is to make use of the approach attributes in SNOMED CT (Access, Surgical approach and Using access device) to match to axis-5 (approach) in ICD-10-PCS. This will lead to better maps when the SNOMED CT concept includes approach information. For example, Laparoscopic adrenalectomy in SNOMED CT has attribute Use access device = Laparoscope, which can be matched to Percutaneous Endoscopic Approach in ICD-10-PCS. Furthermore, the mismatch of excision (removal of part of a body part) and resection (removal of all of a body part) between SNOMED CT and ICD-10-PCS can potentially be resolved by clues in the name of the SNOMED CT concept. For example, SNOMED CT procedures with ‘complete’, ‘total’, ‘entire’ or ‘radical’ in their names can be mapped only to resection, while those with ‘partial’, ‘segment’, ‘wedge’, ‘biopsy’ or the pattern ‘excision of x of y’ (e.g., Excision of cyst of vulva) can be mapped only to excision. In the subset that we reviewed, many of the incorrect mappings can be avoided in this way.

Combining the two maps offers an additional way to use the maps. Maps that agree exactly are much more likely to be correct (93%) than each map alone. The overall performance of the ontological map is just fair because only 45% of the mappings are potentially useful (correct or broader maps). If we restrict to the subset of the ontological map that is corroborated by the lexical map (same map target or agreeing at the 3-character level), the proportion of useful maps will rise to 60%.

In future, we will explore ways to improve the ontological map by curating the body part matches and making use of other ICD-10-PCS axes and SNOMED CT defining attributes. We will also study the possibility of first mapping ICD-9-CM procedure codes to SNOMED CT, and then using the General Equivalence Maps (GEM) published by CMS1 to map from ICD-9-CM to ICD-10-PCS. The ICD-9-CM volume 3 index is another potential resource that can enhance lexical mapping, since the ICD-9-CM index is more comprehensive than the ICD-10-PCS index.

We recognize the following limitations in our study. We only focused on surgical operations but the scope of ICD-10-PCS includes other types of medical procedures (e.g., obstetrical procedures, imaging studies). For the ontological map, we only studied excision procedures, and the results may not be generalizable to other types of operations.

**Conclusion**

We explored two automatic ways of mapping SNOMED CT surgical procedures to ICD-10-PCS: lexical mapping of the ICD-10-PCS index, and ontological mapping by aligning SNOMED CT attributes and ICD-10-PCS axes. Lexical mapping yielded 89% of useful maps (correct or broader) but the coverage of SNOMED CT was limited. Ontological mapping had higher coverage but only 45% of the maps were useful. The performance of the ontological maps can potentially be improved by refining the body part matches and making use of additional information in the SNOMED CT names and defining attributes.

**Acknowledgements**

This work was supported in part by the Intramural Research Program of the National Institutes of Health and the National Library of Medicine.

**References**

Integrating Process Mining and Cognitive Analysis to Study EHR Workflow

Stephanie K. Furniss, MS¹, Matthew M. Burton, MD¹², Adela Grando, PhD¹, David W. Larson, MD, MBA³, David R. Kaufman, PhD¹

¹Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ; ²Office of Information and Knowledge Management and ³Department of Colon and Rectal Surgery, Mayo Clinic, Rochester, MN

Abstract

There are numerous methods to study workflow. However, few produce the kinds of in-depth analyses needed to understand EHR-mediated workflow. Here we investigated variations in clinicians’ EHR workflow by integrating quantitative analysis of patterns of users’ EHR-interactions with in-depth qualitative analysis of user performance. We characterized 6 clinicians’ patterns of information-gathering using a sequential process-mining approach. The analysis revealed 519 different screen transition patterns performed across 1569 patient cases. No one pattern was followed for more than 10% of patient cases, the 15 most frequent patterns accounted for over half of patient cases (53%), and 27% of cases exhibited unique patterns. By triangulating quantitative and qualitative analyses, we found that participants’ EHR-interactive behavior was associated with their routine processes, patient case complexity, and EHR default settings. The proposed approach has significant potential to inform resource allocation for observation and training. In-depth observations helped us to explain variation across users.

Introduction

Health information technologies (HIT), such as electronic healthcare records (EHRs), are expected to bring significant advancements to healthcare delivery through improved management and availability of patient information. Thus far, there have been mixed results from HIT implementation and use. Problems include EHRs not integrating smoothly into clinical work processes and impacting workflow as seen in altered sequences in which tasks are performed, the duration required to complete tasks, the allocation of tasks among workers, development of workarounds, some resulting in adverse events that compromise patient safety and quality of care delivered. The absence of a focus on system usability and on understanding patterns of workflow is a major impediment to adoption and widespread use.

Usability studies typically employ user-satisfaction surveys, focus groups, expert inspections and experiments involving usability testing. Although these methods are informative, they involve a reliance on subjective judgment, may lack reliability and do not provide a sufficiently rich window into the clinical workflow process. Alternatively, there have been numerous studies of workflow that vary in method and scope. There is a need to scrutinize EHR workflow in situ to surface patterns of interaction, characterize the distributions of those patterns and elucidate the factors that underlie them. In this study, we integrate a quantitative process mining analysis of sequential patterns of data access with a qualitative analysis of user performance to investigate and explain clinicians’ work processes. Specifically, we focus on EHR workflow associated with a routine information gathering task (InfoGather).

Background

There is ample evidence to suggest that the implementation of HIT can negatively impact clinical workflows and thereby create staff dissatisfaction, inefficiency and HIT-mediated errors. Current technologies place a burden on clinicians’ working memory and increase cognitive load, which is associated with medical errors and risks to patient safety. Cognitive load reflects the demands on user’s working memory, and is a function of task complexity, user’s skill level, and system usability. The productive use of HIT is partly dependent on the degree to which it can provide cognitive support for tasks that comprise clinical workflow. It is also reasonable to assume that experienced practitioners can develop efficient and effective methods for executing routine tasks—such as information gathering, progress note documentation and order entry—that better leverage the affordances provided by the EHR. We can also hypothesize that clinicians employ suboptimal strategies that result in unnecessarily complex and inefficient trajectories that are more time consuming and error prone. These patterns are empirically discoverable through automated computational approaches that identify patterns of interaction. Further, we recognize that EHR workflow is not performed in vacuum, but rather is connected to a web of actions, interactions, relationships and dependencies between clinicians and work components (e.g., patient, clinician, information, tools, etc.). This necessitates convergent
methods to surface the various factors that shape interaction. The regularities of cognitive work can only be discovered through detailed, time-intensive study of the specific setting\(^9\).

We have developed a methodological framework that draws on three research traditions\(^8,\ 10\): cognitive engineering, distribution cognition and computational ethnography. Each framework provides a theoretical lens, identifying important foci and a set of methods that illuminate different facets of workflow. The cognitive engineering approach focuses on both the usability of the system or interface in question and in the analysis of users’ skills and knowledge\(^11\). In analyzing performance, the focus is on cognitive functions such as attention, perception, memory, comprehension, problem solving, and decision making. The approach has a lengthy history in the study of human-computer interaction in general\(^11\) and in its application to EHRs\(^12\). The cognitive engineering approach has also been used to explain why users employ suboptimal or inefficient procedures or strategies in interacting with systems\(^13\). The theory of distributed cognition (DCog)\(^9\), conceptualizes cognition as distributed across people and artifacts, and dependent on knowledge in both internal (e.g., memory) and external (e.g., visual displays, paper notes) representations as well as their interactions\(^14\). One can employ DCog to characterize workflow as the sequence, or propagation of internal and external representational states across media, settings and time\(^9\).

Computational Ethnography is an emerging set of methods for conducting human-computer interaction studies\(^5\). It combines the richness of ethnographical methods with the advantages of automated computational approaches. Zheng and colleagues define computational ethnography as “a family of computational methods that leverages computer or sensor-based technologies to unobtrusively or nearly unobtrusively record end users’ routine, in situ activities in health or healthcare related domains for studies of interest to human–computer interaction.” Sequential pattern analysis employs log files to search for recurring patterns within a large number of event sequences. The analysis can be used effectively in combination with other forms of data such as ethnography or video-capture of end-users performing clinical tasks. Zheng et al\(^15\) investigated users’ interaction with an EHR by uncovering hidden navigational patterns in EHR logfile data. Various patterns were seen to be at variance from optimal pathways as suggested by designers and individuals in clinical management. Similarly, Kannampallil et al\(^16\) used workflow logfile data to compare the information-seeking strategies of clinicians in critical care settings. Specifically, they characterized how distributed information was searched, retrieved and used during clinical workflow.

In a previous feasibility study, we conducted a process mining analysis with manually-curated event log data from (Morae™) video recordings\(^13\). We found patient case complexity was associated with the complexity of the clinician-EHR interactive behavior for the computer-based pre-rounds information gathering task. Two analyses were conducted. The first characterized the most common patterns of screen transitions. The second analysis quantified the frequency of each screen transition pattern. We observed 27 total screen-transition patterns, each employed 2 to 7 times. We also correlated patterns with interaction measures including mouse clicks and task duration. The objective was to characterize the difference in complexity for each pattern. We observed that, on average, a screen transition resulted in 2 to 2.5 mouse clicks. The task durations per patient were highly variable and may be associated with other factors such as variation in clinical case complexity.

The objectives of this study are to explain variation in EHR workflow by integrating quantitative analysis of empirical patterns with an in-depth qualitative analysis of user performance. This study is part of a larger research project in which we seek to characterize, evaluate, diagnose and improve clinicians’ workflow in post-operative hospital care\(^10\).

**Methods**

**Clinical Setting & Participants**

Research was conducted in the Colon & Rectal Surgery Department (CRS) at Mayo Clinic, Rochester, MN, an academic tertiary healthcare center equipped with a comprehensive EHR since 2005. Patient data is accessed through a customized interface, Synthesis. In CRS Rochester, patients are cared for primarily by surgeons, fellows, resident physicians, hospitalists, nurses, and pharmacists. Hospitalist, in this context, refers to nurse practitioners (NPs) and physician assistants (PAs) who have responsibilities similar to a resident physician. This study was centered on the hospitalist or resident physician, who share responsibilities for coordinating across members of the patients care team, delivering direct patient care, order entry and documentation. Surgery residents work for an attending physician’s service for 6-weeks before cycling to their next service. To date, we have observed four hospitalists, a PA (H1) and three NPs (H2, H3 and H4), and two residents, a 2nd year (R1) and 4th year (R2). H1, H2, H3 and H4 were experienced users of the system and routinely performed the tasks we observed. At observation, they had worked in the unit between 2 and 3 years. R1 and R2 were doing a rotation in the unit and were less experienced users of the system.
This work represents an extension of a surgery practice redesign project, which sought to understand clinical processes and information needs to inform design of new technologies that can improve patient safety and quality and efficiency of health care delivery. It was reviewed by the Mayo Clinic Institutional Review Board (IRB) and judged to be exempt as human subjects’ research.

Pre-Rounds Information Gathering Task (InfoGather)

The data were collected as the clinicians were completing pre-rounds information gathering task (InfoGather). In context of workflow, InfoGather occurs close to the start of the day shift, approximately 6:00 am. Hospitalists and residents round together immediately afterwards. The goals of the task are to access the most recent information on patients’ medical status, review care plans, as well as to anticipate patient needs for the current day. It is clinicians’ first task and serves to anchor their understanding of their patients and their workload. To conduct the task, each clinician reviews patient data in the computer and paper-based information resources, and annotates a paper document that is subsequently referenced and modified throughout their shift.

Data Collection: System event log files & Observation

We observed clinicians in context of their routine workflow and collected ethnographic data from an electronic source (i.e., system-generated log files for the observed participants and the primary EHR application (Synthesis) used by participants for the task).

Synthesis is a customized interface developed by the Mayo Clinic Hospital in Rochester, MN for EHR data aggregation and visualization. We retrieved system-generated event log files for six participants for the six-week period that coincided with the residents’ (R1 and R2) rotation in the CRS department. We also retrieved EHR event log files for the four hospitalists for an additional two-week period that coincided with other observations in the department. At minimum, each event (row) in a log file has a User ID (i.e., clinician ID), an Event Description (e.g., “Activated tab: Labs”) and a Time Stamp (with date and time). Events that are associated with a patient chart also have the patient’s clinic number.

EHR event log files record users’ interactions with the EHR interface, to include selection of a patient chart in the Navigation Panel as well as screen tabs and their associated subtabs in a patient’s chart. The Synthesis application window includes a list of patient records in a panel on the left-side of the screen (Navigation Panel). Synthesis includes a number of screens, separated into tabs, for viewing patient data (top of Figure 1). There are a total of 13 tabs to include, Summary, Labs, Medications, Vital Signs, Intake/Output, Document/Images, Assessment/Cares, Allergies/Immunizations, Patient Facts, Clinical Problem List, Orders and Viewers/Reports. Several tabs have subtabs which allow access to other screens. For example, the Labs tab has 7 subtabs to include, Labs, Microbiology, Pathology and Pending Labs. For all participants, the Summary screen is divided into six equally sized sections, each with a predefined subset of patient data for Allergies, Intake/Output, Medications, Documents, Vital Signs, and Labs. For example, only a patient’s lab data from the last 24 hours are shown in the Summary screen. The clinician would go to the Labs tab to see all past lab results.

We employed Morae™ video capture and think-aloud protocol of participants engaging in InfoGather to allow for retrospective task analysis. Morae™ software is used for usability studies and it records user activity with no interruption to the user’s work. The software provides a screen capture (see Figure 1), and allows use of a webcam to capture audio of participants verbalizing their thoughts (think-aloud) as well as video recording of the participant’s face or hands (inset image in the lower right corner in Figure 1).

We have a broader understanding of the context of the work environment because it was the setting for a larger research project in which we also conducted semi-structured interviews of clinicians from varying roles (e.g., hospitalist, senior resident, and nurse), collected artifacts including paper documents, observed clinicians’ work across tasks and reviewed patient charts. Interview questions aimed to reveal details of clinicians’ key clinical work activities, to include purpose or goal, tasks associated with each activity and resources used. Retrospective patient chart review was performed to understand the complexity of a patient’s clinical state.
Sequential Data Analysis

We employed temporal data mining (i.e., process mining) methods to identify clinicians’ patterns of EHR interaction performed to complete InfoGather. The analyses were conducted using a business process mining tool, Disco™ version 1.9.3, a process-mining workbench used for business process management. Process mining has been used for a wide range of purposes in relation to business\(^1\) and for adherence to guidelines in healthcare\(^2\). The input of Disco is a set of event logs (in our case, EHR logfiles associated with CRS clinicians), which can be processed, analyzed and visualized. Logfiles were preprocessed using Python. Code was written to de-identify log files by replacing clinician IDs and patient clinic numbers with a study ID. Video recordings of observed cases were reviewed with associated log files to understand how the Event Descriptions aligned with users’ behavior.

Quantitative descriptors examined in this study include the number of cases per screen pattern, screen frequency and screen transitions, which were derived from the log files. These descriptors allow us to quantify and compare participants’ interactive behavior required for the task. The quantities provide relative measures of work and reveal insights into the usability of EHR tools, patient case complexity and individual clinician’s interactive strategies. We describe and examine the variation across patient cases and individual clinicians. Further, we integrate other methods (e.g., chart review, sequential analysis, qualitative analysis) to better explore the factors that contribute to the variation.

Qualitative Analysis & Case Study

We conducted qualitative analysis of clinicians’ think-aloud verbalizations to explain patterns of EHR workflow. One of our objectives was to investigate the causes of repeat screen viewing. This was accomplished by reviewing video recordings of clinicians performing the task. In addition, a case study was selectively used to present detailed analysis of clinical work. It provides an illustration of observed behavior with qualitative data interwoven with quantitative descriptors to better understand users’ behavior. For the selected case study, patient case complexity was evidenced by its quantitative descriptors; screen transitions, mouse clicks and task duration for the selected patient case were more than twice the clinician’s, H1’s, average across H1’s observed cases. We previously published these quantitative descriptors and sequential analysis for a subset of cases, the 66 observed patient cases across five clinicians\(^{13}\).
Results

Screen transition pattern analysis

To investigate clinicians’ EHR interaction patterns for InfoGather, a routine computer-based task, we applied process mining to EHR-generated event log files. Our sample consisted of 1569 patient cases across 6 clinician participants. Participants accessed and viewed 26 different EHR screens. Among them are 12 of the 13 main display tabs (the seven most viewed are shown in Table 1). Also included is the Navigation panel (N), which is a collapsible vertical panel on the left of the EHR interface. We defined it as a screen in this study because it is relevant to users’ EHR-interaction for accessing patient charts. Navigation was accessed for nearly all patients (99.7%, Table 1) because it includes the patient list and the search field for a user to access a patient chart. Once a patient chart has been opened during the user’s session, the chart can be reopened by selecting the patient in the Navigation panel (N) or by selecting the chart’s tab along the top of the Synthesis screen. Summary (S), Labs (L), and Vital Signs (V) were viewed for more than half of all cases, and Documents/Images (D) and Intake/Output (I) screens were viewed for more than two-thirds of all cases suggesting the importance of these displays as information sources. Seven screens (D, I, S, L, V, VwR and M) were viewed more than once for some cases and up to 7 times (S and L) (see max repetitions in Table 1). Repeat viewing is analyzed in greater detail below.

Table 1. Screen statistics. Case Frequency is the number of cases in which the screen was viewed at least once. Absolute Frequency is the number of times the screen was viewed. Max Repetitions is the highest number of times the screen was viewed per one case.

<table>
<thead>
<tr>
<th>Screen</th>
<th>Screen Symbol</th>
<th>Case Frequency (% total cases)</th>
<th>Absolute Frequency</th>
<th>Max Repetitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation panel</td>
<td>N</td>
<td>1565 (99.7)</td>
<td>1649</td>
<td>3</td>
</tr>
<tr>
<td>Documents/Images</td>
<td>D</td>
<td>1055 (67.2)</td>
<td>1426</td>
<td>6</td>
</tr>
<tr>
<td>Intake/Output</td>
<td>I</td>
<td>1212 (77.2)</td>
<td>1425</td>
<td>5</td>
</tr>
<tr>
<td>Summary</td>
<td>S</td>
<td>870 (55.4)</td>
<td>1171</td>
<td>7</td>
</tr>
<tr>
<td>Labs</td>
<td>L</td>
<td>828 (52.8)</td>
<td>976</td>
<td>7</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>V</td>
<td>836 (53.3)</td>
<td>966</td>
<td>5</td>
</tr>
<tr>
<td>Viewers/Reports</td>
<td>VwR</td>
<td>179 (11.4)</td>
<td>182</td>
<td>2</td>
</tr>
<tr>
<td>Medications</td>
<td>M</td>
<td>140 (8.9)</td>
<td>162</td>
<td>3</td>
</tr>
</tbody>
</table>

There were 519 variants of screen sequence patterns (Patterns in Table 2). The 15 most frequent patterns account for just over half of all cases (52.6%). All patterns start at the Navigation panel (N). Upon selecting a patient in Navigation (N), the user is immediately transferred to a screen in the newly opened patient’s chart. As represented in the 15 patterns shown in Table 2, transitions lead from Navigation to Documents/Images (N-D), Navigation to Summary (N-S), and Navigation to Viewers/Reports (N-VwR). This is because the users had one of these three screens set as the default opening screen. Documents/Images (D) displayed when H2 and H3 opened a patient’s chart, whereas Summary (S) was set as default for H1, R1 and R2 and Viewers/Reports (VwR) was the default for H4. Due to default settings, H1, R1, and R2 navigated through the Summary screen (S) for all of their patients, but observation of their behavior revealed that only R1 used Summary (S) to access patient data for the task. Because Summary (S) is not used by H1 and R2, navigating through this screen is an unnecessary “cost” for these clinicians.

Table 2 also indicates the percent of clinicians’ patient cases for which the clinician followed each pattern (normalized by clinician’s total to reduce the bias of varying sample sizes). The most frequent screen transition pattern occurred for 132 cases: Navigation to Documents/Images to Intake/Output (Pattern 1: N-D-I). It was followed by H2 for 23% of H2’s cases, by H3 for 27% of H3’s cases, and by R1 one time (0.3% of R1’s cases). The second most frequent screen sequence occurred 67 times: Navigation to Viewers/Reports to Intake/Output to Vital Signs to Labs (Pattern 2: N-VwR-I-V-L). It was followed by one provider—H4 for 41% of H4’s cases (Pattern 2). Among the top 15 patterns, three other patterns were each followed by one provider—24% of H1’s cases (Pattern 6: N-S-L-V-I-D), 19% of H2’s cases (Pattern 7: N-D-I-V-L), and 11% of H4’s cases (Pattern 14: N-VwR-I-V-L-D). Due to the screen default settings, H2 and H3 sometimes followed the same patterns, while H1, R1 and R2 occasionally followed the same patterns, and they never followed H4’s patterns. More than half of the remaining cases (418; 26.6% of total cases) exhibited a pattern that appeared only once (Patterns 102-519). We use the number of sequence patterns with only one case associated as the measure of variation. Thus, H3’s task performance had the least variation (18% of H3’s cases had a
pattern that appeared once), whereas R1’s task performance had the most variation (38% of R1’s cases had a pattern that appeared one time) (Table 2).

Table 2. Screen sequence patterns and frequency measures for InfoGather. Screen sequences are shown for the 10 most frequent patterns. Frequency gives the number of patient cases per pattern. The third column expresses the frequency the clinician uses a screen sequence pattern as a percent of their total cases. EHR screen codes: N (Navigation Panel), D (Documents/Images), S (Summary), L (Labs), V (Vital Signs), I (Intake/Output), VwR (Viewer Reports). *The percent of cases in which clinicians had a unique pattern served as a preliminary measure of variation.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Screen Sequence</th>
<th>Frequency (cases/pattern)</th>
<th>Percent of Clinician’s Patterns</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N – D – I</td>
<td>132</td>
<td>H1 0.23 H2 0.27 H3 0 &lt;0.01 H4 0 R1 0 R2 0</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>N – VwR – I – V – L</td>
<td>67</td>
<td>H1 0 0 0 H2 0.41 H3 0 0 H4 0.01 H5 0.12</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>N – S – V – I – D</td>
<td>65</td>
<td>H1 0.13 0 0 0</td>
<td>H2 0.22 0.03</td>
</tr>
<tr>
<td>4</td>
<td>N – S</td>
<td>95</td>
<td>H1 &lt;0.01 0 0 0</td>
<td>H2 0.22 0.03</td>
</tr>
<tr>
<td>5</td>
<td>N – S – V – I</td>
<td>66</td>
<td>H1 0.16 0 0 0</td>
<td>H2 0.02 0.06</td>
</tr>
<tr>
<td>6</td>
<td>N – S – L – V – I – D</td>
<td>68</td>
<td>H1 0.24 0 0 0</td>
<td>H2 0 0 0</td>
</tr>
<tr>
<td>7</td>
<td>N – D – I – V – L</td>
<td>46</td>
<td>H1 0.19 0 0 0</td>
<td>H2 0 0 0</td>
</tr>
<tr>
<td>8</td>
<td>N – D</td>
<td>52</td>
<td>H1 &lt;0.01 0.04 0.12</td>
<td>H2 0.02 0</td>
</tr>
<tr>
<td>9</td>
<td>N – S – D</td>
<td>45</td>
<td>H1 0 0 0 0</td>
<td>H2 0.05 0.13</td>
</tr>
<tr>
<td>10</td>
<td>N – D – I – L</td>
<td>38</td>
<td>H1 0 0.02 0.12</td>
<td>H2 0 0 0</td>
</tr>
<tr>
<td>11-101</td>
<td>91 sequences</td>
<td>2-40 each</td>
<td>H1 - H2 - H3 - H4 - H5 - H6 -</td>
<td>-</td>
</tr>
<tr>
<td>102-519</td>
<td>418 sequences</td>
<td>1 each*</td>
<td>H1 0.21 0.22 0.18 0.27 0.38 0.30</td>
<td>1.56</td>
</tr>
<tr>
<td>Total</td>
<td>(case count)</td>
<td>1569</td>
<td>H1 288 H2 248 H3 274 H4 162 H5 393 H6 204</td>
<td>585</td>
</tr>
</tbody>
</table>

Although some of the complexity can be accounted for by users’ system settings, others may be accounted for by the interface or may reflect provider efficiency. For example, Pattern 6, employed by a single clinician for 24% of the clinician’s cases, involved sequential transitions from left to right corresponding to the order of tabs along the top of the screen. Similarly, a different clinician employed sequential transitions that correspond to the order of tabs from right to left (Pattern 7). Still others may reflect variation in patient case complexity, which can be inferred from observations as discussed in the next section.

Analysis of Repeated Views: Of the 66 observed patient cases, 31 had at least one screen that was viewed two times (1/9 for H1, 7/21 for H2, 11/16 for H3, 10/12 for R1, and 2/8 for R2). For example, in the screen sequence N-D-I-L-D, which H3 followed for three patients, Documents/Images (D) was viewed twice per patient. We inferred reasons for redundant screen viewing from clinicians’ observed behavior. For most of H3’s patients, Documents/Images (D) was viewed twice per patient because it appeared to be the clinician’s routine process. H3 first viewed D when a patient’s chart was first opened because D was set as the default screen for this user. H3 explained, “Whenever I launch a patient, I’m looking at the notes to make sure there was no weird note put in overnight.” H3 would view D, which H2 followed for three patients, twice per patient because it appeared to be the patient’s routine process. H2 first viewed D when a patient’s chart was first opened because D was set as the default screen for this user. H2 explained, “Whenever I launch a patient, I’m looking at the notes to make sure there was no weird note put in overnight.” H3 would view D towards the end of the task as well, which would allow H3 to review the notes in context of what H3 learned about the patient during the task. H2 also had D set as the default screen, but, unlike H3, the default did not appear to be useful to H2 for several cases. Instead, H2 seemed to use a two-phase approach. First, for most cases, H2 exhibited a consistent screen sequence (i.e., N-D-I-V-L) at the start of the task. Then, for some patients, H2 also visited additional screens, perhaps to see if there were things missed. R1 had the highest percentage of cases with redundant screen viewing. This is not surprising because R1, a second-year resident physician, was relatively inexperienced with the EHR and the CRS practice. R1 could not easily synthesize and consolidate information from the EHR. R1 selectively uses screens with representations that can provide better cognitive support. For example, R1 views both Summary (S) and Labs (L) screens consecutively and multiple times per task. R1 stated “the way they do electrolytes [in the tabular form in the Labs screen], I can’t even sort through that in my mind very quickly so I go back to the skeleton here [on the Summary screen].” In this case, the most recent lab values are represented succinctly in fishbone format on the Summary screen and were the preferred representation.

Case Study: Micro-Analysis of Qualitative and Quantitative data
To explain variation in clinicians’ patterns, we drew on in-depth observation and qualitative analysis of clinicians’ think-aloud. Here, we present a detailed task analysis for one patient case observed in H1’s InfoGather workflow. H1’s screen transition pattern was not repeated for any other case (Pattern 113: N-S-L-S-L-V-I-D). A review of the patient’s chart conveyed the clinical complexity of the patient case in the reason for admission, length of stay, surgical procedures, number of medical services involved in care, and discharge requirements. The patient’s hospitalization was a readmission for a leak and infection. A leak is an abnormal break in the wall of an organ, such as the colon, that allows for an abnormal transfer of contents from the organ to another organ or the body cavity. Observation was conducted on the thirteenth day of the patient’s hospitalization, which was the discharge day. During the patient’s stay, the patient underwent a re-operation and CRS consulted three other services to assist in patient care—critical care, pain service and infectious disease. These consultations are indicative of patient complexity and increased communication needs because information was distributed across additional members of the patient’s care team.

To complete InfoGather, H1 reviewed patient information in the EHR and annotated a paper artifact (paper print out of the electronic handoff document) with patient data, tasks and reminders. Table 3 gives H1’s think-aloud verbalizations, EHR screens viewed and running time for the one patient case. The verbalizations revealed patient information that H1 gathered from each screen (screen captured by the Morae™ video recording). For example, as shown in Table 3, blood pressure is read on the Vital Signs screen (time 01:15), and oral intake volume is read on Intake/Output screen (time 02:18). H1’s verbalizations also reveal data gathered from the Pain Service and Infectious Disease Service notes on the Documents/Images screen (D) in the patient’s chart (time 02:55 to 08:22).

**Table 3. Case study narrative.** H1’s think-aloud, screens viewed and running time for one complex patient case.

<table>
<thead>
<tr>
<th>Time (mm:ss)</th>
<th>Screen Viewed</th>
<th>H1 Narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00</td>
<td>Navigation (N)</td>
<td></td>
</tr>
<tr>
<td>00:01</td>
<td>Summary (S)</td>
<td></td>
</tr>
<tr>
<td>00:02</td>
<td>Labs (L)</td>
<td>alright so she has... great.</td>
</tr>
<tr>
<td>00:03</td>
<td>Summary (S)</td>
<td></td>
</tr>
<tr>
<td>01:08</td>
<td>Labs (L)</td>
<td>She’s a mess. I'm thinking I'd like to hear everything going on [with this patient] because my electronic service list can only tell me so much.</td>
</tr>
<tr>
<td>01:15</td>
<td>Vital Signs (V)</td>
<td>It looks like she's a little hypotensive so I go all the way back to the beginning, which is only 24 hours. I'm trying to go back to see her admit blood pressure so that if I get called about her blood pressure today, at least I'll be familiar if she came with low blood pressure. Alright, I feel better. And again, it's oral intake. I ignore the intermittent infusions. I ignore tubal ligations. number 1 drain nothing. number 2 drain..</td>
</tr>
<tr>
<td>02:18</td>
<td>Intake/Output (I)</td>
<td>Okay, this is good; I need to know this. This is the discharge planning note cause she will go home with IV antibiotics. So I need to make sure _____ for her. So I write down kind of what I need. IV meds, _____, _____ care... per agency protocol. Ordering IV antibiotics is very difficult, I mean outpatient, when I'm setting them up for outpatient. Because all of these things have to be there before they dismiss but we're not supposed to write things ahead of time so it gets kind of hard. recommending lab work and dc PICC at end of therapy. Okay. So then what I do, since IV antibiotics I have to find infectious disease [note]. So Zosyn 3.375 q six. PO qd through the 25th...continuous infusion...13.5</td>
</tr>
<tr>
<td>02:55</td>
<td>Documents/Images (D)</td>
<td>Okay, this is good; I need to know this. This is the discharge planning note cause she will go home with IV antibiotics. So I need to make sure _____ for her. So I write down kind of what I need. IV meds, _____, _____ care... per agency protocol. Ordering IV antibiotics is very difficult, I mean outpatient, when I'm setting them up for outpatient. Because all of these things have to be there before they dismiss but we're not supposed to write things ahead of time so it gets kind of hard. recommending lab work and dc PICC at end of therapy. Okay. So then what I do, since IV antibiotics I have to find infectious disease [note]. So Zosyn 3.375 q six. PO qd through the 25th...continuous infusion...13.5</td>
</tr>
<tr>
<td>05:58</td>
<td></td>
<td>...400 3d. Let's see what pain service is wanting. Because she is an involved patient, that's why I'm looking at this stuff. ... _____, _____, _____...Tylonol and Topomax... So I'm looking at her Sinogram. It says [drain #1] should be flushed daily with 10cc of saline. So then I have to go into MICS</td>
</tr>
<tr>
<td>08:22</td>
<td>MICS: Home Screen</td>
<td>[I] click on [Surgeon] patient, [then] her name, [then] Inpatient Order entry to see if it was done.</td>
</tr>
</tbody>
</table>
Discussion

We characterized clinicians’ EHR-interactions for the pre-rounds information gathering task (InfoGather) by applying process-mining methods to EHR-generated event log files. We hypothesized that there would be a few screen patterns that could explain a majority of the cases because it is a relatively simple task, there are not many screens that have primary patient data, and clinicians may have preferred patterns of screen transitions, which they follow for most of their patients. Screen viewing patterns may be motivated by an intent to seek out new information (e.g., new lab results). An alternative hypothesis is that there is a large amount of variation in EHR-interactive patterns, which may be explained by the differences in patient problems and patient states. Patients with similar problems and profiles (e.g., age and comorbidities) may require the same information-gathering strategy for the task.

There were 519 variant screen sequence patterns to describe the 1569 cases. No pattern described more than 10% of the sample. Fifteen patterns (3%) accounted for just over half of all cases. Because a majority of cases can be described by 3% of total screen patterns, it may suggest that EHR systems can be designed to better facilitate information gathering across screens. On the other hand, there were 418 patterns that only occurred for 1 patient case, which suggests there is much task variation across users or patient types. No pattern of three or more screens was followed by more than 3 of the 6 clinicians (Pattern 8, N-D, employing 2 screens was followed by 4 of 6 participants). Observations helped to explain some of the screen pattern variation. Across the 6 participants, there were 3 different settings that determined the first screen that was displayed when a patient chart was opened; H1, R1, and R2 had Summary (S) set as default, H2 and H3 had Documents/Images (D) set as default, and H4 had Viewers/Reports (Vwr) set as default. The different screen default settings caused variation in screen patterns.

We hypothesized that the clinicians with less expertise would have more variation in screen transitions. The 6-clinician participant pool limits our ability to address this issue, but it can seed hypotheses for further testing. We expected R1, the least experienced user of the EHR among study participants as well as the less expert clinician, to have the most pattern variation. We observed that 38% of R1’s cases exhibited a pattern that occurred only once compared to a 22% average variation across the four hospitalists. More variation is indicative of an incomplete mental model (e.g., understanding of where needed patient information is located or knowledge of potential shortcuts to access data). This is consistent with research showing that more experienced users develop robust mental models21. When a user follows the same pattern for many patients, it may be indicative of the user’s spatial mental model of the system—where information is distributed across sources (applications and their screens). It may also reflect the user’s information needs, and their preferred information sources. R2, in relation to R1, had more clinical expertise and was a more experienced user of the EHR. As expected, R2’s pattern variation was lower than R1’s (30% versus 38%). In relation to the hospitalists, we defined R2 as having equal or more clinical expertise but a less experienced user of the EHR in the CRS department. As expected, R2’s pattern variation was also greater than the hospitalists’ (30% versus 22% average).

Event log file analysis revealed that residents viewed more and different screens, whereas the hospitalists viewed fewer unique screens during the task. This may be additional evidence that residents are less experienced users of the system and, consequently, not as certain where to find patient data. Alternatively, it could be evidence that residents and hospitalists differ in terms of their information needs (e.g., residents are monitoring different or additional patient care goals so they need different or additional patient data). Further evidence that hospitalists and residents differ in terms of information needs and task goals came from R2’s think-aloud, in which the resident appeared to do more decision-making on patients’ care plans than the hospitalists did.

We hypothesized that most clinicians would conduct the task similarly because the routine nature of the task and, though there may be some variation in care goals across patients, there were care goals common to all patients in the unit. There are any number of reasons that may explain variation in these cases including each clinician’s idiosyncratic strategies and patient-case differences. A second-order analysis in which the sequence was not an exact match (e.g., different starting point, but otherwise follows the same pattern) may reveal additional similarities not detected by the first-order analysis. For example, the number of variant patterns is in part due to the variation in chart default settings. By looking at similarities in smaller units of screen sequences, it may be more informative about shared processes and information needs across clinicians.

We triangulated quantitative variables with patient chart review and qualitative data and found clinicians’ EHR-interactive behavior was associated with their routine processes, patient case complexity, variant screen sequence patterns, and EHR default settings. We presented a case study to help convey these findings. In particular, this case
study exemplifies non-routine requirements of the InfoGather task. Our data suggests the selected patient case is complex because the quantitative descriptors (reported in 13) are more than twice H1’s average interactivity across all nine observed patients. Data presented in this study further supports this assertion: 1) H1 described the patient as “an involved patient”. 2) To complete InfoGather, H1 followed a screen transition pattern variant that was unique in our sample. The pattern involved 7 transitions between main-tab EHR screens and 12 transitions including the EHR’s subtabs and other clinical applications. 3) Among the 7 EHR screens, H1 visited two screens twice (S and L). This may suggest that redundant screen viewing for a patient case is also an indicator of patient case complexity. On the other hand, redundant screen viewing may suggest users’ inefficient task performance, an indicator of system usability or of task complexity (e.g., conflicting data gathered from another screen or from the paper handoff document may cause the clinician to return to a previously viewed screen). As an outlier case, it surfaces some complexities of completing the task and negotiating the information sources.

Process mining enables researchers to variably focus data analysis on clinicians, patients, time, tasks and interactions, thereby providing insights into different dimensions of workflow. The level of analysis is limited only by the granularity of the event logs. A limitation of our data collection is that we only looked at event logs from one clinical information system and in one setting. Future work will combine event logs generated from additional clinical information systems. As we examine how automated event log files are useful to study EHR workflow, we can also inform what user behavior is captured in systems’ event log files. That is, we can determine if there are particular user-computer activities that event logs could capture and that would be informative of users’ cognitive work.

We conducted qualitative analysis on the subset of observed cases to investigate why some screens were viewed two or more times (redundant screen viewing) for one patient case. Three clinicians repeated screen viewing for a third or more of their patient cases. Reasons varied from clinician’s lack of experience, lack of cognitive support provided by the system to synthesize and consolidate patient findings, system default settings and clinicians’ deliberate routine work process. It was assumed that default setting improved workflow so that clinicians would not have to visit the screen twice. It could be that the default setting serves them better on another task. Future design could support task-based navigation through the system, which could reduce cognitive workload.

Future research will employ regression analysis to determine if there are associations between measures of patient complexity (e.g., procedure, primary diagnosis, number of medication at admission, days in hospital, number of services involved in care, etc.) and measures of task complexity (e.g., number of screen transitions, task duration, etc.). The case study supports further investigation of this. If there is an association between patient case complexity and task complexity, it could be used to quantify increased workload on clinicians from complex patients. Also, it could be evidence toward development of advanced clinical decision support systems that facilitate team awareness and collaboration of complex patients.

Cognitive studies grounded in DCog framework examine clinical work in the context of actual practice and can identify issues in human and system performance. A contribution of our approach to cognitive research is that we leverage large data sets of system-generated event log files to investigate users’ behavior in conjunction with observational methods to explain variation in the empirical findings. In this study, we demonstrated the value of integrating quantitative data analysis with qualitative data to examine EHR workflow. This study indicates that process mining techniques can be used to evaluate variation of clinicians’ task behavior across many clinicians, which can potentially be used to direct resource allocation for observation or training when patterns of interaction seem aberrant or inconsistent with clinical pathways. This study presents a part of a larger methodological framework that we are developing for the study of clinical and EHR workflow. In the larger research project, we are studying workflow from multiple perspectives (e.g., tasks, clinicians, patients, tool). As we branch beyond analysis of a single task, we are exploring how to use system log files to examine care team coordination activities, particularly to study patient-centered clinical workflow. Variation in team processes surfaced through process mining can be used to focus observation efforts.

**Conclusion**

The presented study approach addresses the need for an integrated, in-depth approach that facilitates broad investigation of workflow across many settings, clinicians and patient cases, while also facilitating detailed analysis. We used system-generated event log files to characterize clinicians’ EHR-interaction patterns for a routine computer-based task, along with observation to explain variation in the patterns. As demonstrated in this study, computational ethnography can be integrated with observation to balance the advantages and limitations of individual data collection methods and to enable collection of a broad and rich data set for studying clinical work.
Acknowledgements

We would like to thank the Mayo Clinic Office of Information and Knowledge Management (OIKM) for funding this research initiative and Research Fellowship for Stephanie Furniss’s doctoral work. The work was partially supported by a Mayo Clinic Professional Service Award to David Kaufman. A special thanks to the clinicians in Colon & Rectal Surgery Division who graciously volunteered to participate in this study. We are grateful to the efforts of Robert Sunday and Katherine Wright who assisted in data collection.

References

Understanding the Knowledge Gap Experienced by U.S. Safety Net Patients in Teleretinal Screening

Sheba M. George, PhD¹,², Erin Moran Hayes, PhD³, Allison Fish, PhD⁴, Lauren Patty Daskivich, MD, MSHS⁵, Omolola I. Ogunyemi, PhD¹,²
¹Charles R. Drew University, Los Angeles, CA; ²UCLA, Los Angeles, CA; ³L.A. Pierce College, Woodland Hills, CA; ⁴Indiana University, Bloomington, Indiana; ⁵Los Angeles County Department of Health Services, Los Angeles, CA.

Abstract

Safety-net patients’ socioeconomic barriers interact with limited digital and health literacies to produce a “knowledge gap” that impacts the delivery of healthcare via telehealth technologies. Six focus groups (2 African-American and 4 Latino) were conducted with patients who received teleretinal screening in a U.S. urban safety-net setting. Focus groups were analyzed using a modified grounded theory methodology. Findings indicate that patients’ knowledge gap is primarily produced at three points during the delivery of care: (1) exacerbation of patients’ pre-existing personal barriers in the clinical setting; (2) encounters with technology during screening; and (3) lack of follow up after the visit. This knowledge gap produces confusion, potentially limiting patients’ perceptions of care and their ability to manage their own care. It may be ameliorated through delivery of patient education focused on both disease pathology and specific role of telehealth technologies in disease management.

Introduction

Diabetes is estimated to affect 21.3 million adults in the United States, with 1.7 million new cases of diabetes diagnosed every year.¹ Compared with non-Hispanic whites, racial and ethnic minorities are more likely to be diagnosed with diabetes, including 9.0% of Asian Americans, 12.8% of Hispanics, and 13.2% of non-Hispanic blacks.¹ California has one of the highest rates of diabetes in the country, with 9.9% of the population diagnosed with the chronic disease.² One complication of diabetes mellitus is diabetic retinopathy, which causes damage to the blood vessels of the retina.³ Diabetic retinopathy is the leading cause of blindness in the United States with more than 100,000 new cases identified each year.⁵

Safety net clinics in the United States offer primary health care services to over 16 million patients nationwide and 2.3 million patients in California, whether or not these patients have the ability to pay for health care services.⁵ Patients in South Los Angeles experience limited access to care, and many are African American and Latino. Limited access to appropriate eye care and delays in diagnosis and treatment can result in advanced disease, such as diabetic retinopathy and vision loss among such populations.³,⁶ In a study of Los Angeles inner city minority residents, patients were 3.5 times more likely to present with advanced diabetic retinopathy and more than 5 times more likely to require immediate intervention, referral or follow-up than newly presenting patients in a predominately white non-urban setting⁶.⁸ The American Diabetes Association (ADA) guidelines for best practices necessitate annual retinal screening examinations, timely disease diagnosis, and on-going treatment to prevent the loss of eyesight due diabetic retinopathy. At-risk patients can benefit from laser photocoagulation surgery if retinopathy is detected early (in an annual examination).⁹

There is a growing body of research evaluating the efficacy of telehealth programs¹⁰,¹¹ and the use of telemedicine to screen for diabetic retinopathy has shown promise as a way to provide screenings to patients in areas where there is a lack appropriate access to these services.¹²¹⁴ Given the shortage of specialty eye care in medically underserved areas, such as South Los Angeles, the larger study of which this paper is a part, examined the use of teleretinal screening as an innovative way to address this gap in care. In the clinical use of teleretinal screening, technology is conceptualized as facilitating communication between a primary care provider and specialists (though other health care team members may be involved); the patient’s retinal images are instrumental to that communication but no direct engagement of the patient by clinicians occurs beyond the generation of retinal image data. Studies of the clinical use of teleretinal screening have focused primarily on providers’ use and perceptions of technology. In this paper, we shift the focus from provider to patient to explore one aspect of the qualitative findings of the larger study: patients’ reported feelings of fear and confusion about teleretinal screening at critical points during the delivery of care. This is especially important to understand given that even with the introduction of teleretinal screening, safety
net screening rates lag behind national screening rates. In order to understand this outcome, it is necessary to explore patient perceptions of telemedicine and the quality of care they received. As the primary setting in which many of the most disenfranchised healthcare users encounter health technologies, safety-net clinics present a particularly important site to understand how structural inequalities preconfigure patients’ interactions with such technologies in the healthcare setting, and how these technological encounters may shape patients’ beliefs and behaviors around the self-management of their chronic disease. Specifically, our findings indicate that these patients’ limited health literacy compounds with pre-existing socioeconomic barriers and low technological literacies to produce a “knowledge gap” that limits patients’ understanding of what is happening with the screening and their ability to comply with prevention goals and treatment plans.

Materials and Methods

Study Setting

The study took place from August 2010 to September 2011 and involved six safety-net primary care clinics in South Los Angeles. During the study period, a total of 2,732 unique patients were screened for diabetic retinopathy by three ophthalmologist readers, with 1,035 receiving a recommendation for referral to specialty care. The study focused on both the technical challenges of implementing and screening patients, as well as the patient and staff perceptions of telemedicine. Each clinic used digital retinal cameras to conduct retinal screenings of Type 2 diabetes mellitus patients. The six clinics primarily serve immigrant Latino and African-American patients in communities that have 28% of the population living below the federal poverty level.

The gold standard for detecting and diagnosing diabetic retinopathy is seven-field 35-mm stereoscopic color fundus photographs and grading protocols, as defined by the Early Treatment Diabetic Retinopathy Study (ETDRS). However, ETDRS photography is impractical for use in clinical settings so the more common method of diabetic retinopathy screening is via an in-person examination with a licensed eye care provider (Optometrist or Ophthalmologist). Diabetic patients in Los Angeles who visit safety net primary care clinics were traditionally referred to the Los Angeles Department of Health Services outpatient eye clinics for yearly retinal screening examinations and treatment of diabetic eye complications, resulting in long wait times for appointments. Prior to the introduction of teleretinal screening, primary care clinic staff report that patients could wait up to eight months to receive an initial retinal examination, diagnosis, and treatment. Complications such as proliferative diabetic retinopathy may result in permanent vision loss when screening and treatment are delayed.

Teleretinal screening for diabetic retinopathy, however, allows for routine screening via retinal images taken with a fundus camera by ancillary staff in primary care settings, with subsequent analysis by trained readers to determine presence and extent of disease. To conduct a teleretinal screening, clinic staff took six retinal images and two external images of a patient’s eyes. The patient’s case, consisting of images and basic biometric data, were then uploaded into EyePACS, teleretinal screening software platform developed at UC Berkley and already in use in over 360 California safety net clinics. Three board-certified ophthalmologists contracted as image readers, assessed patients’ cases, recommended referrals for further care, and rated quality of the retinal images provided.

Approach and Procedures

Focus group techniques were used to assess the acceptability of teleretinal screening among Latino and African American patients who had received screening at one of the participating clinics. Six focus groups were conducted with two groups in English with patients who self-identified as African American and four groups in Spanish with Latinos, most of whom spoke Spanish as a primary language. The subjects were 18 years or older, diagnosed with diabetes, both male and female, separated by gender. Each focus group included 6-9 patients, with the total sample size for the focus groups equaling 42 participants (29 Latinos, 12 African American, 1 unreported). Our study sample is representative of the Los Angeles county health services planning area in which the clinics are located as well as the study clinics’ overall populations. First, 9.5% of LA County population was diagnosed with diabetes in 2011, perfectly mirroring the diabetes rates in the six clinics from which our retrospective study sample.

<table>
<thead>
<tr>
<th>Table 1: Racial/ethnic breakdown of patients in a) Six Study clinics, b) Teleretinal Screening Study sample and c) Focus groups sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Six Study clinics</strong></td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>African American</td>
</tr>
<tr>
<td>Latino</td>
</tr>
<tr>
<td>Asian</td>
</tr>
</tbody>
</table>

Our study sample is representative of the Los Angeles county health services planning area 6 in which the clinics are located as well as the study clinics’ overall populations. First, 9.5% of LA County population was diagnosed with diabetes in 2011, perfectly mirroring the diabetes rates in the six clinics from which our retrospective study sample.
was selected in the same time period. Our sample for the study represented 20% of the diabetes population in the six study clinics and for the most part, reflects the overall clinics’ patient populations. For example, Table 1 shows that the racial ethnic breakdown of the patient populations at the six study clinics and our study samples were very similar. Because recruitment of participants for the qualitative study was opened up to screened African American and Latino patients from all participating clinics, we believe that both our larger study sample and focus group study sample are reflective of the clinics’ overall patient pool. A script guided interviews, with questions arranged by category, focusing on patient perceptions about accessibility, acceptability, and satisfaction with the teleretinal screening that they received, as well as barriers to compliance with ophthalmologic referrals. Each focus group lasted approximately 90-120 minutes and the monolingual Latino groups were conducted in Spanish. All groups were audio taped and transcribed by a professional transcription company (to which participants consented). Upon each subject’s completion of participation in the focus group session, the subject was provided with remuneration of $50.

**Data Analysis**

Using Atlas ti software to help manage and analyze the data, focus group transcripts were coded and indexed by team members to develop analytical categories based on qualitatively informed and modified grounded theory techniques of analysis. Team members performed open coding independently to identify themes and to generate codes. Then transcripts were recoded with the team-developed codes. Constant comparison within and across categories allowed researchers to check codes against the rest of the data to establish categories that reflect the nuances of the data, key themes and theoretical insights. Scientific rigor is strengthened through use of common procedural guidelines for qualitative studies. Credibility of the results is supported through use of data from 6 focus groups with carefully chosen participants, and a team with diverse research expertise and backgrounds. An iterative mode of data analysis by multiple team members increased dependability of the findings. Transferability of findings is made possible through published description of the methods and findings.

**Results**

**Demographic Data**

Table 2 shows participant characteristics. 29% of focus group participants were African American and 71% of participants were Latinos. 50% of all participants were female, with both African American and Latino focus groups being split evenly between males and females. 60% of Latinos had not completed high school, while on average African Americans had attained a higher education status with 50% having attended some college. 62% of participants earned less than the 2012 poverty level of $11,170 per year. Computer ownership was greater among African Americans (50%) than among Hispanic participants (30%). Latinas (56%) reported a slightly higher rate of computer ownership than Latinos (44%), while only 33% of African American females reported owning a computer compared to (67%) of African American males. Among Latinos, internet use was low, with only 17% of participants reporting access.

**Focus Group Results**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>African American N(%)</th>
<th>Latino N(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (mean)</td>
<td>56</td>
<td>48</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>6 (50)</td>
<td>15 (50)</td>
</tr>
<tr>
<td>Female</td>
<td>6 (50)</td>
<td>15 (50)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College grad</td>
<td>2 (17)</td>
<td>3 (10)</td>
</tr>
<tr>
<td>Some college</td>
<td>6 (50)</td>
<td>3 (10)</td>
</tr>
<tr>
<td>Secondary grad</td>
<td>1 (8)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Some secondary</td>
<td>2 (17)</td>
<td>18 (60)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (8)</td>
<td>5 (17)</td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-500</td>
<td>4 (33)</td>
<td>11 (37)</td>
</tr>
<tr>
<td>501-999</td>
<td>3 (25)</td>
<td>11 (37)</td>
</tr>
<tr>
<td>1000+</td>
<td>5 (42)</td>
<td>8 (26)</td>
</tr>
<tr>
<td>Own computer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>4 (67)</td>
<td>4 (44)</td>
</tr>
<tr>
<td>Female</td>
<td>2 (33)</td>
<td>5 (56)</td>
</tr>
<tr>
<td>Internet access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5 (42)</td>
<td>5 (17)</td>
</tr>
<tr>
<td>Female</td>
<td>3 (60)</td>
<td>2 (40)</td>
</tr>
</tbody>
</table>

Table 2: Participant Characteristics
Findings indicate that in teleretinal screening, the patients’ knowledge gap becomes consequential at three critical points: (1) when patients’ pre-existing personal (structural) barriers to care became exacerbated in the clinical setting; (2) during the encounter with technology during the screening; and (3) during the doctor-patient follow-up after the screening. Summary of results are presented in Table 3 below.

1) Patients’ Personal Barriers (Structural barriers)

Patients in this study group expressed concerns about their a) personal barriers to care, including the ability to access care, the quality of care that they receive, the costs of care, and the ability to follow treatment plans as prescribed. Patients also expressed b) confusion regarding the pathology of their disease.

1.a) Nature of personal barriers to care: Participants generally report a variety of barriers to accessing healthcare including, obtaining appointments, long wait-times, need for multiple appointments, loss of patient information by clinic, lack of clarity regarding the cost-of-care to patient, and the receipt of unexpected bills after care. These barriers were compounded by the fact that many participants indicated time pressures due to work and family obligations, transportation issues, as well as, financial pressures. For example, many patients spoke of the difficulty of complying with treatment plans due to daily schedules. Work regularly intervened in participants’ ability to take prescribed medications at the proper time, to eat healthily and consistently throughout the day, and to find the time to exercise. One Latino respondent put it this way, “I sell food in the morning and I practically don’t eat. I eat at around 3:00 pm and that’s not good, so they say…That’s bad, I have to be eating a little every now and then, and eat only once a day.” Compliance with treatment plans was especially difficult for African American men, who spoke of this difficulty at twice the rate of other respondents across ethnicity and sex. One respondent reported that his experience of incarceration made complying with prescribed diet extremely difficult. He said, “I was incarcerated, I cheated (on my diet plan) because we have common food. So when I eat ice cream, I’ve got to work a little harder to burn it off… the doctor told me, mine comes from bad diet. Okay? It’s not hereditary, it’s bad diet.”

1.b) Confusion regarding disease pathology: Further complicating this situation, patients expressed confusion about pathology of their disease. Patients uniformly expressed concerns related to unresolved symptoms they attributed to their diabetic status. Such concerns included, dizziness, exhaustion, circulation-related foot pains, and, most significantly for this study, vision problems. One female Latina participant expressed confusion about her disease pathology in this way: “I had a question - I’m hearing terms like diabetic retinopathy…And then I’ve heard macular regeneration [sic]…And then [the doctor] talked about cataracts…Are those three different ailments of diabetic…symptoms?” Like this patient, the majority of participants expressed uncertainty about how these health conditions. Patients indicated difficulty communicating these concerns to their providers, particularly due to the medical language and communication styles of the clinical setting. Significantly, patients felt that they were not given adequate education before, during, or after their appointments to address their concerns about their health.

<p>| Table 3: Three primary points at which the “knowledge gap” is produced in teleretinal screening |</p>
<table>
<thead>
<tr>
<th>1) Patients’ Personal Barriers (Structural Barriers)</th>
<th>2) Encounter with Technology During Screening</th>
<th>3) Impact of Lack of Clinician Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nature of personal barriers to care</td>
<td>2. Role of the retinal screening</td>
<td>3. Confusion over lack of follow up</td>
</tr>
<tr>
<td>Lack of resources (time, money)</td>
<td>Confusion about role of teleretinal screening in disease management</td>
<td>Patients were confused about:</td>
</tr>
<tr>
<td>Lack of information</td>
<td>Misunderstanding the teleretinal screening as a visual acuity test</td>
<td>- What to do after screening</td>
</tr>
<tr>
<td>Limited ability to negotiate clinical interactions</td>
<td></td>
<td>- How to get results (most patients never received results)</td>
</tr>
<tr>
<td>Inability to follow treatment plan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Confusion regarding disease pathology</td>
<td>2. Concerns about the technology during screening</td>
<td>3. Quality of patient-provider interactions</td>
</tr>
<tr>
<td>Confusion about pathology of diabetes</td>
<td>Lack of confidence in operators abilities</td>
<td>Some patients felt that the providers:</td>
</tr>
<tr>
<td>Difficulty communicating concerns with providers</td>
<td>Confusion about how photos are shared between clinic and specialist</td>
<td>- Were rushed</td>
</tr>
<tr>
<td></td>
<td>Fear of technology</td>
<td>- Didn’t take time to answer questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Were not thorough enough</td>
</tr>
</tbody>
</table>
Despite this limitation, patients generally expressed a desire to know more about their disease and learn how to manage it better, but were uncertain how to obtain this knowledge easily.

2) Encounter with Technology during Screening

With regards to diabetic retinopathy screening, patients demonstrated a similar lack of clarity as to the technical aspects of the teleretinal screening process specifically, and retinal screening generally. Despite having received teleretinal screening in-clinic, patients were uncertain about 1) the role of the retinal screening in their disease management and 2) what was happening during their encounter with technology in the screening.

2.a) Role of the Retinal Screening: Patients across all the focus groups misunderstood the purpose of the retinal diabetic retinopathy screen. Most patients did not understand that retinal screening is used to detect signs of diabetic retinopathy, with some patients confusing it with visual acuity testing. For example, one Latina female expressed her confusion about the role of teleretinal screening in preventing her diabetes-related retinopathy as follows: “I've been wanting to ask you, the test that we had done for our eyes. Well, they already told me that I had liquid and it was –

retinopathy, with some patients confusing it with visual acuity testing. For example, one Latina female expressed her confusion about the role of teleretinal screening in preventing her diabetes-related retinopathy as follows: “I've been wanting to ask you, the test that we had done for our eyes. Well, they already told me that I had liquid and it was leaking…does that have anything to do with, if I need to increase my [glasses] lenses? The doctor didn’t say exactly the name of the problem. He only said you have this problem, your eyes are starting to leak from inside and that’s it. [I don’t] understand the word.” One African American woman put the confusion this way: “I've been a diabetic for, like, 24 years, and I just started feeling like sometimes I can’t read street signs…When I left from the clinic [after the teleretinal screening], they referred me over to Martin Luther King…got me a new set of glasses and everything…now I’m good, you know. Just I’ve got a large nerve in the back of my eye, and that’s about it.” Some patients who were screened shared that they continued to experience poor vision or undefined eye pain, symptoms although their screening had been negative for diabetic retinopathy. These patients were concerned and felt that the screening received might have been inadequate since they continued to exhibit impaired visual acuity. The examples above demonstrate that the patients did not understand their pathology and the purpose of screening.

2.b) Concerns about the Technology during Screening: A majority of patients were nervous about their encounter with teleretinal diabetic retinopathy screening technologies. Patients expressed concerns about potential human error in handling of private health information over the Internet, mishandling of the technology by clinic staff, or the feeling that something could go wrong during the screening that might result in damage to the eyes (due to the drops, flash, or camera hitting the eye.). One Latina woman put her fears this way, “[I felt] afraid because I have a sister who lives in Honduras and she had an eye operation and instead of improving her vision they damaged her… And that’s why I got scared… because [my sister] couldn’t see but they made it worse…”

Patients expressed a variety of emotions about the absence of a specialist in the room during the exam, from indifference to resignation to confusion. African Americans in particular seemed to feel resigned to the lack of a specialist in the room. For those who understood the role of the specialist, not seeing a specialist directly seemed to be consistent with their experiences of other types of healthcare in safety net clinics. While the responses were not wholly negative, African American women respondents reported preferring to have the specialist in the room but understanding that this was not always possible due to lack of resources in the clinic. For example, one African American woman said, “…I like personal attention and personal care, so of course, I want the best hands possible. I would want to be in the hands of an ophthalmologist or an optometrist, but in this case, I was going to get a screening. It’s kind of like how you have the PA’s versus a doctor who do the vital signs or whatever, I just thought this was a routine sort of thing. I didn’t know it was like new age…. Well, had I been told that an online person would be evaluating the screening; it may have made me feel like it was less personal. But, to be honest with you, knowing that [the specialist is remote] doesn’t make me feel that different then.” In other words, this patient population is accustomed to being seen by lower level staff than the specialist, or even the primary care doctor.

Having retinal images taken by a technician (medical assistant) rather than a specialist, however, led some individuals, particularly African American and Latino men, to feel a sense of insecurity in the ability of the technicians,. These patients felt that medical assistant photographers lacked the necessary medical/technical knowledge to handle the equipment and could not answer health-related questions. One African American patient described the medical assistants as ‘fresh out of school’ while another said, “[T]he people that they have in there [operating the screening equipment], they’re kids. They don’t know anything. You know? They're looking at you – they're no specialists and they look like they just come out of a training school.” A Latino male elaborated, “…I don’t know if they were well trained or they needed more training to learn how to work with the machine because the person that was taking the photos…wasn’t sure on how to work the machine and I think what that person was supposed to do was something simple and yet it took longer due to the lack of experience.”
3) Impact of Lack of Clinician Follow-up

The large majority of patients in these focus groups expressed concern and confusion about the follow up procedures and the regular lack of follow up, post-screening. In addition, although many patients were pleased with their providers, some patients felt concern about the quality of the interactions they had with the clinician.

3.a) Confusion over lack of follow up: The majority of patients expressed confusion about what to do following the teleretinal diabetic retinopathy screening and about how to get their results, how suggested diagnoses would be returned, and what procedures typically follow the test. Most patients reported never receiving the results, post-screening. One Latino male participant said, “I, at least, expect to know my results because knowing that I have diabetes, you tend to have a little fear…[so] they should tell [me] how I’m doing inside.” While standard protocol is that patients should be informed of a negative screening result, patients were not informed that it may happen at the next follow up visit with the primary care provider and not via special outreach. Another African American male put his concerns over the lack of results this way. “[T]he first time I was diagnosed was with another doctor, and she sent me to the [screening] and I had it. That was over a year ago. I remember [her] just getting back to me with having the tests done again, but I’ve never had anyone respond to me from the clinic about the results of the tests that I had there.” Since patients did not understand the standard procedure for a negative result, they felt they lacked quality care. Part of their confusion stemmed from the challenges of sharing results and information internally among the clinics and between clinics and specialists, which occasionally resulted in patient screening results being lost or misplaced. One Latino respondent described being asked to locate his own results and bring them in to his doctor himself. He said, “They mailed me mine because when I went to see the doctor, the doctor asked me for them and they told me they were going to fax it here but the doctor said he didn’t have them and he told me I had to bring them in. So I had to lose a day from work to pick them up and bring them over to the clinic.” Other patients did not understand why they had not heard back from their doctor after the teleretinal screening.

3.b) Quality of patient-provider interactions: Some of the patients felt that their doctors and medical assistants were rushed and did not take enough time to answer all their questions, and as a result were not being thorough. One Latino male put his concern this way, “…I asked what [the screening] was for and the [medical assistant] didn’t know how to answer and then I asked the doctor and he said, ‘No, everything came out fine.’ But that was the only thing they said.” Similarly, an African American male felt that the doctors were rushed and didn’t take the time to educate him. He said, “…your doctor doesn’t say anything, he just tells you, ‘I’m going to give you some drops’, you know, and send you on your way. But he doesn’t sit down plainly [and tell you what to do].” Consequently, patients such as these were concerned that the care providers’ inability to explain the teleretinal screening was indicative of an inability to provide patients with on-going, quality care.

Discussion

Telemedicine has been lauded in the health information technology (HIT) literature as a means of increasing access to specialty care in low resource settings, such as rural or inner city community clinics. The Chronic Care Model (CCM) includes the use of HIT as one avenue for addressing the challenges of chronic disease care management through use of clinical information systems and clinical decision support21. Health communications studies have also demonstrated the ways in which HITs can be used to a) support individual health information seeking22, b) facilitate doctor-patient communication23, and c) promote consumer understanding about prevention and treatment opportunities24. Overall, patients who participated in this study, too, shared the general perspective that telemedicine advances the delivery of care by improving communication between providers. While this is likely the case in terms of the general adoption of telemedicine at sites where no specialty services existed before, this research team has reported elsewhere the challenges of implementing effective workflow communication pathways in the delivery of telemedicine workflows25. Challenges include missing or misreported patient information, primary care providers’ difficulty interpreting diagnostic results to patients, and the inability to track patients across multiple information management systems. Although unknown to patients directly, these clinical workflow difficulties are experienced by patients in ways not obviously related to the communication challenges in the clinic as reported in the results above.

Patients in the populations studied here face a number of challenges in accessing healthcare. In addition to challenges reported in the literature, participants in this study describe several others, which we conceptualize as a knowledge gap in the clinical setting, which appears to be produced at three points during the delivery of care: (1) through exacerbation of patients’ pre-existing personal barriers in the clinical setting; (2) through encounters with technology during screening; and (3) and through lack of follow up after the visit. In the following discussion, we elaborate on the patients’ experiences of the knowledge gap at the three points and consider the types of knowledge/information that patients are lacking and potential approaches to fill these gaps in the teleretinal screening process.
**Addressing patients’ pre-existing barriers: The role of health literacy**

Participants in this study lacked adequate understanding regarding the purpose, processes and goals of preventative tests such as the retinopathy screening. This finding reflects larger trends in the literature that indicate low levels of health literacy around diabetes pathology among medically underserved minority populations. Retinopathy screening is complicated by the fact that for the most part, the disease is asymptomatic until it gets to the advanced stages, so patients without symptoms may be baffled by the need for screening while patients with non-diabetic symptoms are frustrated by the screenings not addressing their existing symptoms. When told that they needed retinal screening, they were not educated sufficiently about the need for screening in the context of diabetes. From the patients’ perspectives, as demonstrated by focus group data, they often believed they were receiving a visual acuity test, rather than a retinal screen related to their diabetes. While these patients had access to diabetes management classes in some of the clinics, it was clear from their comments that they had not been provided any consistent, structured education about the role of retinal screening in diabetes. Based on our interviews, clinic providers were aware of patient concerns and many actually attempted to provide information to patients but these efforts were neither comprehensive nor systematic and thus obviously not effective. It is important to acknowledge the difficulty that these providers faced in describing disease progression to patients with limited literacy in their native languages. Nevertheless, a comprehensive educational approach that systematically addresses differing levels of patient health literacy with sensitivity to their preexisting barriers is one key factor necessary in addressing the knowledge gap among safety net patients going through teleretinal diabetic retinopathy screening.

Patients also faced significant preexisting socioeconomic barriers that shaped their access to care and their understanding about the importance of care. For example, 25% of the African Americans and 63% of the Latinos had received only a secondary education with, 60% of the Latinos having only some secondary education. Furthermore, 58% of the African American participants and 74% of the Latino participants made less than $1000 per month. Lack of resources such as time, money and information and limited ability to negotiate clinical related interactions resulted in great challenges for these patients to follow recommended treatment plans to manage their health conditions, such as diabetes. It is in the context of such severe preexisting barriers that their knowledge gap becomes consequential for follow through on teleretinal screening. These patients were challenged by limited health literacy when receiving health care in general, and in this instance, resulting in misunderstandings about their diabetic disease pathology and progression. This was most especially observable when patients misconstrued generalized symptoms of discomfort related to the eye (including symptoms unrelated to diabetes or diabetic retinopathy), to have been caused by their disease.

**Addressing patients’ fears during screening: The role of technological literacy**

A second point at which the knowledge gap is consequential is during the visit is when the patient encounters the technology associated with teleretinal screening. Patient engagement can be difficult to attain in the safety net because the typical patient may not understand the health technologies or the implications of their use. Whereas patients from a wide range of backgrounds are increasingly using computers and accessing the internet for health-related reasons, a gap continues to exist between those who can more effectively access and use information technology compared to those who do so less effectively. Healthcare scholarship has linked this gap to a pre-existing “digital divide” among certain groups of HIT users. The low rate at which patients owned computers and had access to the Internet among participants in our study highlights the digital divide experienced by these populations. Only 30% of the Latino participants owned a computer and 17% had access to Internet compared to 50% of the African Americans who owned a computer and 42% had access to the Internet. Whereas the digital divide has conventionally been conceptualized in terms of individual and population-level lack of access to technology (computers, internet), health communications scholarship has noted that ‘access’, particularly in the context of healthcare delivery, also includes meaningful use of technologies and exploring the social context in which people’s interactions with HIT take place. Extending the conceptual reach of the ‘digital divide’ beyond the scope of patients’ individual abilities or technology savvy to include technological literacy in the clinical setting sheds light on the role that limited literacy plays in individual experiences of access in other contexts, especially in the safety-net setting.

Patients who lack familiarity with medical technologies are more fearful of the equipment and are anxious that the machine may damage their eyes during the exam. This perception may discourage patients from complying with screening appointments. Furthermore, patients were concerned that photographers were medically and technically under-skilled. This may have been caused, in part, by high turnover among medical assistants, as well as minimal on-going training and interaction with the technology to maintain photographer technical skills. Also, in one clinic...
out of the six involved, clinic protocols prevented medical assistants from offering any medical or technical education during the screening. For some patients, photographer lack of confidence or unwillingness/inability to answer questions about care coupled with the absence of a specialist evoked concerns about the quality of care and the adequacy of the screening.

Patients’ lack of familiarity with process of teleretinal screening and with medical technologies could be addressed by informing patients about how teleretinal screening works. The first step may be to assess patients’ past experiences with retinal screening since this is the starting point for educating them about teleretinal screening. For example, it was not clear from our focus group data as to how many of our participants had ever received retinal screening for diabetic disease management. Patients may differ on their prior experiences and thus may need different types of information about teleretinal screening. Participants reported being put at ease with the technology and the screening process when there were interactions with their providers, particularly when information and education was provided during the exam. Thus, a systematic approach to educating patients about teleretinal screening should include a focus on a) how technology is used in the screening process, b) how training is provided to photographers/medical assistants and other clinic personnel in the appropriate use of the technology, c) how security of confidential electronic health data is managed during data transfer and d) the role of telehealth technologies in disease management. Such a systematic educational approach would improve patient technological literacy, allay some of their fears and begin to address the knowledge gap in teleretinal screening.

Addressing patients’ confusion after the visit: The role of Improved Provider-Patient communication

Following the exam, most patients appear to have little understanding of the significance of their diagnosis. Typically, patients were not contacted regarding a negative teleretinal screening result, whereas during an in-person retinal screen, a patient is given negative results immediately. Patients do not know how to interpret the lack of follow up by clinics; rather than an indication of a negative screen, lack of communication is perceived as no result at all. As a consequence, patients do not understand their own health-status and are unable to make effective use of the teleretinal screening results to direct and manage their own care.

Since patients do not adequately understand the pathology of diabetes and diabetic retinopathy, they have difficulty interpreting the results that may be delivered by the primary care provider (i.e. one patient reported that the provider only said “You’re fine”). Providers, who must concern themselves with all aspects of a patient’s pathology, may not take adequate time to explain the significance of a negative retinal screening resulting from diabetic retinopathy to a patient or try to work around patients’ perceived limitations. From the patient perspective, the lack of patient education about the role of teleretinal screening in their disease management plan seems to reinforce patient perception that, despite having received the screening, they are receiving subpar or incomplete care. For patients who already face a number of challenges to accessing care, this perception may hinder their ability or desire to comply with preventative screening in the future.

The confusion experienced by these patients with regards to post-screening results seems aligned with the lack of standardized education or consistent communication from providers. Low health literacy patients should be given clear information about posttest protocols - what it means to get no call, that negative results are good and that they have the right to call and find out screening results. Providers and other clinic staff may need to consider how to best communicate information about screening results with adequate detail and depth, nuanced to address the knowledge gaps of patients with differing levels of health and technological literacies and socioeconomic barriers.

Conclusion

In summary, patients’ pre-existing socioeconomic barriers along with their level of understanding regarding their disease and teleretinal screening technologies, combine to produce a knowledge gap in the clinical setting. Confusion about their disease, compounded by fear and lack of knowledge about disease management technology, may cause patients to feel powerless to understand or address their health concerns, their personal health/disease status, and result in unsatisfactory encounters with providers. Each of these outcomes has the potential to discourage patient compliance with treatment plans and to diminish patients’ ability to manage their own health care.

These findings support other studies regarding the challenges of health care access and health literacy in this population, which has been well documented over the last ten years \[1,2,26,27\]. Our research builds on these findings by presenting new information about how patients understand and perceive the digitization of clinical communication and the use of new technologies as an aspect of their chronic diabetes management. Despite their specific confusions related to teleretinal screening, most participants were generally pleased with the quality of care they receive at these community clinics. Patients also reported a strongly positive perception of the incorporation of new technologies...
into the primary care visit, particularly in the context of eye health. They associated such new technologies with progress in science and improved quality of care. While patients were generally satisfied with the overall care they received from their primary care providers, patients were concerned about specific aspects of teleretinal screening that arise as a result of the technology and the clinic protocols.

These findings suggest that patients in this clinical population, who exhibit both limited health literacy and limited technological literacy, need to be educated before, during, and after their telematic appointments. Patient education regarding both the health and technological aspects of the screening is essential in encounters with medical staff of all levels. Such education helps to ensure that patients understand the role of technology in their disease management, the procedures of the test, what to expect after the test, and how to interpret the results of their test. This is likely to have an impact upon patient understanding of teleretinal screening and, ultimately, upon patient compliance with diabetic retinal screening protocols. Our findings call for going beyond a definition of the “digital divide” that focuses on the patient’s personal access to and use of technology to consider how their level of technological literacy, health literacy and socioeconomic barriers together affect their experience of the increasing use of HIT in health care institutions. This will be particularly important, as patients are required to engage more actively in self-management of chronic diseases such as diabetes. Patient education will be important to ensure that telemedicine in the clinical setting does not become an additional barrier that contributes to the knowledge gap that exists among these populations.

There are some limitations to our study and findings. We have a relatively small convenience sample and our participants are not statistically representative of the wider population in inner-city settings and this limits the generalizability of our study’s conclusions. However, as is common to qualitative methods, they represent information-rich cases, homogeneously stratified across race and ethnicity, to allow in-depth understanding of their perceptions and experiences of teleretinal screening among these groups. While five years have elapsed between data collection and presentation of these findings, the results presented here are applicable to other resource-poor clinics implementing teleretinal screening programs to serve similar socio-economic populations.

This paper explored the way in which African American and Latino diabetic patients’ with less access to healthcare experienced teleretinal screening as an encounter with HIT in U.S. urban community clinic settings. This study’s significance lies in its identification of how patient disparities in health literacy, technological literacy and access to healthcare interact to produce a knowledge gap in their experience of teleretinal screening, potentially compounding their confusion and fear about the progression and management of their disease. Consequently, in the safety net clinic setting with a low health literate population, without adequate patient education, teleretinal screening may represent a new challenge to effective disease management. Patient education regarding the function and purpose of telehealth technologies, particularly in the specific contexts of their use in the health care setting, and improved provider awareness and communication of this information to the patient is critical to ensuring effective teleretinal screening, improving patient engagement in the self-management of chronic illnesses and addressing the knowledge gap experienced by most such patients in the clinical setting.

Acknowledgments

This project was supported by the NIH under grant number U54 MD007598-01S2 (formerly U54 RR026138-01S2). We would like to thank Dr. Richard Baker for his role in shaping our clinical understanding of teleretinal screening.

References


Address for correspondence
Sheba George, PhD. Center for Biomedical Informatics, Charles R. Drew University of Medicine and Science., 1731 East 120th St., Los Angeles, CA 90059. Email: shebageorge@cdrewu.edu
Foraging for Information in the EHR: The Search for Adherence Related Information by Mental Health Clinicians

Bryan Gibson, DPT, PhD, 1,2 Jorie Butler, PhD,1,3,8 Maryan Zirkle, MD, MS, MA, 4 Kenric Hammond, MD, 5,6 Charlene Weir, PhD1,2

1 IDEAS 20 Center George E Whalen VA Medical Center, Salt Lake City, UT; 2 Department of Biomedical informatics, University of Utah, Salt lake City, UT; 3 Geriatric Research Education and Clinical Center George E. Whalen VA Medical Center, Salt Lake City, UT; 4 Portland VA Medical Center, Portland, OR; 5 Puget Sound Health Care System, Seattle WA; 6 University of Washington, Seattle WA; 8 Division of Geriatrics, University of Utah, Salt Lake City, UT

Abstract

In this project we sought to qualitatively describe clinician’s search for information related to the complex construct of adherence. Nineteen think aloud observations and semi-structured interviews were conducted with mental health providers as they prepared for a patient visit. The transcripts were coded according to constructs from information foraging theory (information goal, patch, scent, enrichment, and opportunity cost). The search strategies uncovered were complicated: provider’s searches were sometimes multi-staged (e.g. a search of the EHR led to further enquiry when interviewing the patient), and involved multiple ‘patches’ (i.e. data from the EHR, the patient and other providers were all sought out). In addition, some information that providers considered relevant to understand adherence related questions was non-obvious (e.g. the absence of specific information was considered a useful cue). Providers’ information search strategies for complex constructs are at times non-intuitive; implications for the design of EHR summarization tools are discussed.

Introduction

Clinicians spend significant time and energy searching the Electronic Health Record (EHR)1. They perform these searches to address clinical questions, and to integrate information into a coherent mental picture of their patient’s status 2. Our ultimate goal is to develop an EHR summarization tool that collects information related to a given clinical question and provides an integrated display. We hypothesize that such a tool would minimizing the time and cognitive resources required of help clinicians to locate and make sense of information relevant to the question at hand.

In service of this larger goal, the purpose of this project was to characterize providers’ information needs and search strategies related to a common, but complex clinical question: is this patient adherent? 3 Specifically, this study focused on the information search of mental health clinicians as they addressed questions related to adherence of their patients with Post-Traumatic Stress Disorder (PTSD). PTSD is a chronic mental health problem that is frequently co-morbid with conditions such as substance abuse and depression 3, and in which adherence to recommended treatment and self-care is critically important 4,5. We chose this use case because it is ideal for exploring clinicians’ search regarding adherence: the chronic and changeable nature of PTSD requires clinicians to repeatedly search the EHR in order to assess their patient’s adherence to recommendations and response to therapy.

Prior Studies Examining EHR Usage

The goal of this study was to examine how clinicians use the EHR to answer clinical questions in a real-world setting. This goal requires an in-depth exploration of both how clinicians frame their questions, and how the EHR is used. Our approach is in line with prior work, which has observed clinicians as they use the EHR 6,7,8,9. The novel component of this work is our focus on clinician’s search process as they use the EHR to address their information needs, based on their mental model of the information space. We believe that this approach can derive important implications for EHR design.
**Information Foraging Theory**

In this study, we used the constructs of Information Foraging theory (IFT) as a conceptual framework. IFT has been used to describe and predict individuals’ search behaviors on the internet\(^\text{10}\), and in bibliographic databases\(^\text{11}\). In addition, IFT and its associated quantitative models have been used to develop and evaluate information search and summarization tools\(^\text{12,13}\). We believe these prior successful applications of the theory suggest its possible utility in the development of tools to summarize the information in the EHR.

Information foraging theory (IFT) is based on an analogy with optimal foraging theory (OFT)\(^\text{14}\), which focuses on the optimization of costs/benefits to an animal in its search for energy from food in a given environment. Pirolli describes the essential idea of information foraging theory as:

"The optimal information forager is the one that best solves the problem of maximizing the rate of valuable information gained per unit cost, given the constraints of the task environment" \(^\text{14}\) p. 8.

The theory proposes that the individual spends their time either “within-patch” foraging, or in one of two types of “between-patch” activities: using proximal cues to make decisions about the potential value of information patches (Scent-following), or molding the information environment to either reduce between-patch foraging costs (Enrichment activities). Table 1 describes the theory’s central constructs and their definitions.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Definition / Examples from the EHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Goal</td>
<td>The information sought relevant to the current task / e.g. remind myself of this patient’s plan of care.</td>
</tr>
<tr>
<td>Patches</td>
<td>Physical or virtual areas of concentrated information yield / e.g. notes, tabs, windows, etc.)</td>
</tr>
<tr>
<td>Scent</td>
<td>The individual’s imperfect perception of the values, cost or access path of information sources obtained from proximal cues / e.g. perceptions of the value of information underlying note titles, tab titles, alerts, etc.</td>
</tr>
<tr>
<td>Resource Costs</td>
<td>The actual costs incurred by pursuing a given information source / e.g. time required, clicks required to locate desired information.</td>
</tr>
<tr>
<td>Opportunity Costs</td>
<td>The benefits that would have been accrued in pursuing a different information source, but were not gained due to the given pursuit instead</td>
</tr>
<tr>
<td>Information Diet</td>
<td>The range of foraging choices the individual makes amongst several potential sources.</td>
</tr>
</tbody>
</table>

Two assumptions of the model should be emphasized. First, the theory assumes that the value of information is not intrinsic but is dynamic and task specific. This assumption has face validity in our proposed use case: clinicians searching the EHR for data related to a patient’s adherence to treatment will naturally weigh certain information sources more heavily than others. Second, the theory does not assume that the individual is classically rational with perfect information and infinite computational resources. Rather it suggests that individuals engage in “satisficing” which involves optimizing within the constraints of imperfect information and limited time and cognitive resources. Therefore, the solutions reached are not assumed to be globally optimal but allow for the specification of “information niches” that may be local maxima. Again, this approach has face validity in clinical practice. We do not expect that clinicians exhaust all possible avenues to address an information need. Instead, they make use of the data that is readily available to them to address a question, and often must make decisions in the face of uncertainty.
Methods

Human Subjects

This study was conducted in accordance with World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was reviewed by the Institutional Review Boards of the George E. Whalen Salt Lake City VA Medical Center (SLCVA) and the University of Utah. Informed consent was conducted with both the provider and the patient.

Overview

This was an observation and interview study of clinicians’ information search in a real clinical setting. Clinicians were asked to think-aloud as they prepared to see a scheduled patient and they reviewed the patient’s chart. This study combined direct observation and interviews with a qualitative analysis. The methods chosen for this study aimed to capture clinicians’ information search as closely as possible to their uninterrupted, unobserved real world processes.

Setting

The George E. Whalen Salt Lake City VA Medical Center (SLCVA) is a 101-bed tertiary care center that provides both inpatient and outpatient mental health services. This study took place in an outpatient mental health clinic and targeted patients diagnosed with PTSD.

Description of Participants

Eight mental health clinicians from the SLCVA’s outpatient mental health clinic were recruited to participate in this study. All clinicians were treating individuals with Post Traumatic Stress disorder (PTSD). Of the eight clinicians who participated in this study, seven are psychologists and one is a social worker. Six are female and two male. They ranged from less than one year of clinical experience (a post-doctoral fellow) to more than 20 years of clinical experience. Clinicians were recruited via both in-person requests at staff meetings and follow-up emails. The first author conducted a total of 19 observations (a minimum of 2 and a maximum of 3 observations per provider) in the outpatient Mental Health clinic. To maximize variability each observation was of a unique patient visit (no repeat visits by the same patient were observed). To ensure that adherence was a relevant concern, all appointments were follow-ups.

Procedures

After completing informed consent, mental health clinicians were observed prior to their appointment with patients. As they prepared to see a patient, clinicians were instructed:

“Please review the chart as you normally would and think aloud about what information you are looking for, what information you are finding, and what you are thinking about.”

The duration of the chart review duration was at the discretion of the clinician but was generally brief (5-7 minutes). After completing the chart review, clinicians participated in a semi-structured interview intended to probe for information goals and search strategies for adherence related information (the Appendix contains the interview guide). The interview lasted 10-15 minutes. When the patient arrived for their appointment, the study was explained to them and they then completed informed consent. The patient’s consent simply provided the study team with access to those notes in the patient’s electronic medical record that the clinician had reviewed in preparing for the visit. The think aloud and the interviews were audio recorded and then transcribed by a medical transcriptionist. All individual identifiers were removed from the texts in the process of transcription.
**Data Analysis**

The study team used an iterative process of qualitative analysis using ATLAS Ti@. Each researcher highlighted text they believed to represent information relevant to information search, and/or concepts of adherence. Highlighted areas were reviewed discussed by the group, and given what Patton refers to as “pre-codes” - short descriptions of the content \(^9\). Many of these pre-codes referred to adherence concepts, patient’s response to therapy, risk assessments, and clinician’s mental models of the information space. The goal was to capture concepts relating to clinician’s mental models of adherence, and the information search processes related to those mental models. After additional discussion and review, the constructs were organized through consensus into higher-level categories according to Information Foraging Theory.

**Results**

The results presented here are organized around the constructs of Information Foraging Theory. The analysis is interpretative, using theory to explain and understand results.

**Information Goals**

Table 2 presents examples of text that were coded as exemplars of information goals. Information goals associated with adherence varied, and included: determining compliance with assigned homework, determining the patient’s response to therapy, identifying if the patient was ready for a specific therapy, and looking for evidence of how “activated” the patient was in terms of treatment. In many cases information goals consisted of discrete pieces of information that would be straightforward to electronically query for and represent in a summarization. However, in several cases during their chart review several clinicians reported developing questions that required further follow-up with the patient. This type of information goal, one in which an original search precipitates a second information search via a different “patch” (e.g. the patient) has implications for EHR design which we will discuss later.

**Table 2. Example Text for Information Goals Relevant to Adherence**

<table>
<thead>
<tr>
<th>Information Goals</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is new with the patient?</td>
<td>“He was supposed to go to event name. So I want to ask him how that</td>
</tr>
<tr>
<td></td>
<td>went for him.”</td>
</tr>
<tr>
<td></td>
<td>“Like I need to know what she’s avoiding, what she’s been</td>
</tr>
<tr>
<td></td>
<td>experiencing, what else is she avoiding”</td>
</tr>
<tr>
<td>Who else is seeing the patient?</td>
<td>“Looking for an initial assessment by his prescriber…but it’s finally</td>
</tr>
<tr>
<td></td>
<td>notated in here who that person is”</td>
</tr>
<tr>
<td></td>
<td>“He presented for PTSD assessment and I want to know who did the</td>
</tr>
<tr>
<td></td>
<td>assessment”</td>
</tr>
<tr>
<td>How is the patient doing? Is he improving</td>
<td>“Some violent urges. Has noted that PTSD therapy is helping him,</td>
</tr>
<tr>
<td>or getting worse?</td>
<td>good.”</td>
</tr>
<tr>
<td></td>
<td>“Has he seen any improvement or any side effects?”</td>
</tr>
<tr>
<td></td>
<td>“He doesn’t have any suicidal thoughts or any homicidal ideation tells</td>
</tr>
<tr>
<td></td>
<td>me that he’s doing well”</td>
</tr>
<tr>
<td>Is the patient going along with treatment?</td>
<td>“I wanted to go check what meds he was prescribed (by a provider</td>
</tr>
<tr>
<td></td>
<td>outside the VA) and I want to ask him about all of these today”</td>
</tr>
<tr>
<td></td>
<td>“I’m going to go to my last note and confirming treatment clinic,</td>
</tr>
<tr>
<td></td>
<td>she’s going to do PE, motivated”</td>
</tr>
<tr>
<td>What is the plan of care?</td>
<td>“We sort of reviewed therapy goals and he wanted to focus on</td>
</tr>
<tr>
<td></td>
<td>decreasing anger at home and focusing on behavioral activation, and</td>
</tr>
<tr>
<td></td>
<td>exploring more leisure activities in his life.”</td>
</tr>
<tr>
<td></td>
<td>“Looking at the plan to see if there’s other stuff in there I need to</td>
</tr>
<tr>
<td></td>
<td>follow up on”</td>
</tr>
<tr>
<td>Is this patient appropriate for therapy X?</td>
<td>“I need to know this because if she’s taking a benzo, that’s really</td>
</tr>
<tr>
<td></td>
<td>important to know for PE (prolonged exposure therapy).”</td>
</tr>
<tr>
<td></td>
<td>“Why else was he referred for a motivational interview?”</td>
</tr>
</tbody>
</table>
**Information Scent**

Table 3 presents text snippets that were coded as exemplars of Information Scent. In most cases these references are simple, explicit cues about informativeness. However, clinicians sometimes mentioned the absence of information as a cue. For example, one clinician noted the absence of emergency department notes as a cue that her patient might be doing better. Several participants mentioned the absence of any notation that the patient was a poor historian as evidence that the patient must be a good historian and several also mentioned that the absence of notes about non-compliance to therapy was a sign that the patient was likely adherent. This particular form of Information Scent, the absence of particular information, has implications for the design of an EHR summarization tool.

**Table 3. Example Text for Information Scent**

<table>
<thead>
<tr>
<th>Scent subtype</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Risk Cue</td>
<td>“Noticed this patient has a high risk for suicide flag”</td>
</tr>
<tr>
<td></td>
<td>“Recipient of Bronze Star so potentially real PTSD stuff.”</td>
</tr>
<tr>
<td></td>
<td>“The behavioral flags and things”</td>
</tr>
<tr>
<td>Adherence Related Cues</td>
<td>“He’s compliant with his hearing aids”</td>
</tr>
<tr>
<td></td>
<td>“She’s motivated. She doesn’t have no-shows.”</td>
</tr>
<tr>
<td></td>
<td>“The no-shows, that’s something I look for especially to see how compliant they have been or avoidant they have been.”</td>
</tr>
<tr>
<td></td>
<td>“. . . Frequency of his appointments is usually a good sign for him. When he’s coming in really regularly it means he’s engaged in treatment.”</td>
</tr>
<tr>
<td></td>
<td>“Regularly filling his medications.”</td>
</tr>
<tr>
<td></td>
<td>“Vets who just fire providers all the time”</td>
</tr>
<tr>
<td>Absence of Information as a cue</td>
<td>“Has he had any emergency department visits? No, good.”</td>
</tr>
<tr>
<td></td>
<td>“There’s nothing noted that he’s not compliant with his current meds.”</td>
</tr>
<tr>
<td></td>
<td>“Poor historian we usually comment on that…”</td>
</tr>
</tbody>
</table>

**Information Patch**

An information patch is the actual location of information (rather than a pointer to that source). Table 4 presents text snippets that were coded as exemplars of Information Patches. In most cases, our participants visited a limited number of patches- mental health notes and more frequently their own notes, were the most common patches searched. However, in several cases clinicians mentioned information locations outside of the EHR (e.g. patients, other clinicians) as important sources. Again this finding has implication for the design of an EHR summarization tool.
Table 4. Example Text for Information Patch

<table>
<thead>
<tr>
<th>Patch subtype</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note Titles and Sections</td>
<td>“Suicide prevention team follow-up note, 21-day contact”</td>
</tr>
<tr>
<td></td>
<td>“Medical notes would become more relevant if they’re doing UAs (Urinalysis)”</td>
</tr>
<tr>
<td></td>
<td>“Most recent prescriber note”</td>
</tr>
<tr>
<td></td>
<td>“There is a part in that states that he is compliant with his medications”</td>
</tr>
<tr>
<td>Assessment Scores (Health Factors)</td>
<td>“One thing that I look at is where his PCL and BDI [standardized disease severity scales] scores are.”</td>
</tr>
<tr>
<td></td>
<td>“All right, we would definitely look at her PCL”</td>
</tr>
<tr>
<td>Other providers as a patch</td>
<td>“Prescriber in the hallway who asked, “Is he taking any meds?””</td>
</tr>
<tr>
<td></td>
<td>“I usually have to go ask the prescriber. It’s not generally in the notes”</td>
</tr>
<tr>
<td>Patients as patch</td>
<td>“Check in with him every time on the homework that I assign from the last session.”</td>
</tr>
<tr>
<td></td>
<td>“I specifically ask questions to try to get at that [patient’s adherence]”</td>
</tr>
</tbody>
</table>

Resource Cost

Resource cost refers to the time or effort needed to find information. Table 5 presents texts snippets that were coded as exemplars of Resource Cost.

Table 5. Example Text for Resource Cost

<table>
<thead>
<tr>
<th>Resource cost subtype</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time/Effort required</td>
<td>“I have such a hard time when they’re on inpatient with substance abuse, finding like a relevant, important note.”</td>
</tr>
<tr>
<td></td>
<td>“We spent 45 minutes in the chart looking”</td>
</tr>
</tbody>
</table>

Enrichment

Table 6 presents texts snippets that were coded as exemplars of enrichment (an individual improving their information environment to improve future yield). Interestingly, several clinicians mentioned a form of enrichment that is unique to the EHR: clinicians often place other clinicians as cosigners on their notes to promote a shared awareness of their patients’ disease state and treatment plans. In addition, several clinicians spontaneously lamented the absence of search tools within the EHR. We note that no text was coded as representative of enrichment by improving within-patch yields, we will return to this in the discussion.

Table 6. Example Text for Enrichment

<table>
<thead>
<tr>
<th>Enrichment subtype</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduce between patch foraging</td>
<td>“I just usually sort it (clinical notes) by title and then I would just look for year, any mental health notes or primary care behavioral health notes”</td>
</tr>
<tr>
<td>Identify Note Cosigner</td>
<td>“I cosign his treatment clinicians over there on my note to let them know how he’s doing.”</td>
</tr>
<tr>
<td>Absence of Search Tools</td>
<td>“There’s no way to search [The notes in the EHR]”</td>
</tr>
<tr>
<td></td>
<td>“Unless there’s like a very particular key word, there’s no way to search”</td>
</tr>
</tbody>
</table>
Discussion

In this project we explored mental health clinician’s information search as they addressed questions related to the adherence of their patients. Our findings indicate that clinicians’ information search in the EHR is complex and at times non-intuitive. For example: clinician’s engaged in information search in one source which precipitated a second search via a different source (e.g. reading a note reminds the provider to ask the patient a particular question during the visit), clinicians reported the absence of information as an important cue (e.g. clinician noting the absence of emergency department notes as a cue that her patient might be doing better) and clinicians communicating with each other by placing each other as cosigners on their notes (a form of enrichment that is unique to the EHR). We believe that these findings regarding searches that are distributed over time and across patches, the absence of information as a “scent”, and the social form of enrichment exemplified by cosigning, have significant implications for the design of EHR search and summarization tools.

This study is one of only a few to use Information Foraging Theory to describe the information search of healthcare workers. Dwairy, et al. studied primary care physicians when addressing clinical questions and found that colleagues and books were preferred sources of information (patches), they concluded that this was likely due to the decreased time required to answer the question compared to an electronic database search 16. Kannampallil, et al. used a think aloud protocol and observation to examine the information search of physicians in an ICU setting. They found that physicians sought out different sources depending upon their specific information goals, and that answering clinical questions required integration of data from multiple sources. In addition these researchers estimated the relative information gain from various sources and found that electronic sources provided more unique information per unit time than paper records 17. Our findings echo this prior work in that the “information diet” of the clinicians we studied included both human and electronic information sources. Our findings expand upon this prior work by identifying the specific patches, scents and enrichment strategies used by clinicians to address their questions regarding their patient’s adherence and by noting areas where our observations will impact the design of our planned EHR summarization tool.

Implications for EHR Design

We believe that our findings offer insights into potential improvements in EHR design. First, this work illustrates how information needs are context dependent, what would seem on the surface to be the same information need: adherence, varies in content and structure depending on why the information is needed. We believe this suggests that EHR designers need to understand the reasons why clinicians might be interested in a specific class of information and then use that understanding to develop an appropriate information display. For example, a clinician wanting to understand their patient’s “dose” of exposure to medical intervention might use the patient’s attendance at medical appointments as one measure but would likely want to correlate this with other measures (e.g. frequency of appointments, number of unique providers, number of medications, etc.). Conversely, a clinician wanting to understand how “activated” the patient is might correlate the same appointment attendance information with a different set of measures (e.g. patient reported activation, reports from other providers, measures of health related behaviors). We believe that our findings suggest that EHRs should provide integrative displays that address the user’s information needs in context. Second, our finding that clinicians use multiple information patches to address a single overarching information need suggests the need for the inclusion of a broader range of data into the EHR. For example including patients’ functional and social data in the EHR 18 would reduce clinicians need to search outside of the EHR for this data, allowing for easier mental integration of this data with clinical data and therefore greater efficiency. Third, the concept of information “scent” seems critical to improving EHR design. Clinicians clearly sought out specific patches depending upon the limited scents available (e.g. note titles). The EHR could provide stronger “scents” regarding the location of specific information types. These scents might include cues as to the content of specific patches as well as critical metadata such as certainty (e.g. provisional vs. confirmed diagnosis) and currency (e.g. historical problems vs. current problems). We hypothesize that an EHR that provided these cues would improve the efficiency and accuracy of clinician’s searches. Finally, our results suggest that clinicians clearly want EHRs that allow them to enrich their information space. This would include: the ability to aggregate data elements that are relevant to specific clinical questions from a variety of patches, the ability to search within patches to more efficiently locate specific data, and tools to communicate and collaborate with other clinicians.
The results of this study will inform our future work in developing an EHR summarization tool. The non-obvious nature of providers’ mental models of the information space suggests that creating a map between clinical questions and information needs will likely require involving providers themselves. One potential method to accomplish this would be for providers to use tagging tools to identify data elements and patches that are relevant to specific clinical questions. This mapping could be available for future use by an algorithm as well as for sharing with other providers (e.g. collaborative search). To address providers’ need for the capacity to enrich within-patch yields, we might provide the ability to search for concepts semantically related to a given keyword within notes. Finally, the use of cosigning to promote shared awareness and the adding of amendments to clinical notes as a tool for collaboration, points to the need for communication and collaboration tools in the EHR.

**Strengths**
This study has several strengths. First, we used a well-established model of information search (Information foraging theory) as our conceptual framework. Second, we used two types of qualitative data (think aloud and interviews) to capture different perspectives on clinicians’ information search. Third, we used iterative thematic coding to minimize bias in our findings.

**Limitations**
This study has limitations. Since the sample size was small and the data collection was restricted to mental health clinicians treating individuals with PTSD, the results may not generalize to other domains, therefore further work is needed to validate these preliminary results. It is also worth noting that we did not address or control for how long the clinicians had been seeing the patient; the duration of this relationship may have affected clinicians information search. Future studies should account for this important factor. In addition our measurement of information scent relied on provider’s report of what they considered cues to important information. In future work we could improve upon this by measuring the clinicians’ gaze, this would allow us to note the information cues that providers both consciously attend to (gaze at and comment upon) and implicitly attend to (gaze at without awareness) and then either pursue or do not pursue, thus providing a much richer representation of information scent. Finally, in this study no text from our transcripts was coded as representative of “opportunity costs.” In future work, this might be addressed by prompting participants to reflect on the expected value of the search within a specific patch relative to querying an alternative source (e.g. look in a different part of the EHR, telephone the patient to answer this question).

**Conclusions**
In this study we combined think-aloud and semi-structured interviews to understand mental health clinicians search for adherence related information. We found that clinician’s information search to understand their patients’ adherence is complex, and at times non-intuitive. We believe this foundational work is a useful first step toward the development of EHR summarization tools that will improve the effectiveness and efficiency of clinicians’ information search.

**Appendix: Semi-Structured Interview Guide**

1. What tells you that the patient is doing better or worse?

2. What in the note informs you if the patient is adherent to recommended treatments or not (whatever the treatment is – e.g. medications, therapy, CBT, group)

3. If the patient is not adherent, what information in the notes is most helpful to determine WHY the patient is non-adherent?

4. What information would inform you about the likely consequences (for the patient and others) that are likely for being non-adherent?

5. Where do you look in the notes to find out if prior interventions have been done to address non-adherence and how well they worked?

6. Looking in the chart, would there be any reason to suspect that the patient is not a good historian?
References

“Scared to go to the Hospital”:
Inpatient Experiences with Undesirable Events

Shefali Haldar¹, Alex Filipkowski¹, Sonali R. Mishra¹,
Cory S. Brown¹, Rashmi G. Elera¹, Ari H. Pollack, MD¹,², Wanda Pratt, PhD¹
¹University of Washington, Seattle, WA; ²Seattle Children’s Hospital, Seattle, WA

Abstract
Involving patients in healthcare safety practices has long been an area of priority and importance. However, we still need to understand how patients perceive undesirable events during their hospital stay, and what role patients play in the prevention of these events. To address this gap, we surveyed pediatric inpatients and caregivers to understand their perspectives on undesirable events. By giving them an opportunity to use their own words to describe their experiences, we found a diverse array of undesirable events. Our qualitative analysis revealed four major types of events that patients and caregivers experienced: mismanagement, communication, policy, and lack of care coordination. We also examined the information needs that patients and caregivers experienced during these situations, and learned how they would prefer to receive this information. Based on these results, we provide recommendations for inpatient technologies that could enable patients to identify and prevent such undesirable events.

Introduction
Despite over a decade of national attention to the problem of patient safety, as described in To Err is Human and Crossing the Quality Chasm¹,², a recent study estimates that over 440,000 people die every year from preventable medical errors in U.S. hospitals³. As of 2016, preventable medical errors are considered the third leading cause of death in the United States, behind only heart disease and cancer⁴. Research over the past decade suggests that patients provide a distinct perspective on safety events that is not captured through staff reports and medical record review⁵,⁶. Weissman and colleagues looked at the overlap between undesirable events reported by patients and by physician review of medical records. Their work showed that 23% of 998 sampled patients reported at least one adverse event, compared to 11% of cases identified through medical record review. Moreover, only 12% of the events reported were discovered through both mechanisms⁶. Thus, to identify a comprehensive view of patient safety, we need to understand patients’ perspectives in addition to clinicians’ perspectives. Although several studies have solicited input from patients, these studies use closed-ended questionnaires based on categories of undesirable events that are predetermined and predefined by clinicians and medical researchers⁷,⁸,⁹. Thus, more research is needed to understand patients’ perspectives in their own words¹⁰.

Previous research has shown that patients can play a significant role in a variety of safety-related processes, such as helping healthcare providers confirm an accurate diagnosis, making appropriate treatment decisions, and choosing a good provider that aligns with the patient’s needs and values¹¹. Patients have also been recognized as actors in identifying, preventing, and reporting medical errors. In 2007, Unruh and Pratt described five cases where oncology patients were involved in the identification, prevention, and recovery from medical errors in an outpatient setting¹². Other studies in the inpatient setting have shown that patients notice errors related to their care and bring them to the attention of their providers¹³,¹⁴,¹⁵. However, we know little about what patients’ experiences and perspectives around patient safety are, or how to develop informatics solutions that support patients in their important safeguarding role.

In support of a strong role for patients, evidence from over 50 studies has shown a positive correlation between patient experience (e.g., respect for the preference of patients, inclusion of patients in decision-making, clinicians’ ability to empathize, and provision of information to enable self-care) and patient safety¹⁶. Despite this close relationship between patient experience and safety, existing technology for enhancing the patient experience is not typically designed to support and engage patients as active participants in the detection or prevention of undesirable events. To examine a related area of the role of existing technologies in supporting patient engagement within the hospital environment, Prey et al. found that technologies served a variety of purposes, including the use of entertainment to decrease anxiety and improve rehabilitation treatment, the delivery of general or tailored information to patients and caregivers,
the enrichment of patient-provider communication, and the enhancement of communication and personalized decision support between patients and caregivers\(^7\). However, of the 17 articles included in the review, only two emphasized the technology’s role in improving safety\(^18,19\). Additionally, Greysen et al. conducted a pilot study using a tablet computer-based learning module to help educate patients about safety in the hospital\(^20\). Although these insights help us understand the role of patient engagement technology in the hospital setting, it remains unclear how similar tools, informed by inpatient safety experiences, can effectively support patients in identifying, preventing, and reporting undesirable events.

In this paper, we address these gaps and present the results of a survey asking pediatric inpatients and their caregivers about their prior experiences with undesirable events. Our study provides a rich understanding of these patient and caregiver experiences in their own words. We also identify information needs that participants felt could have prevented these undesirable events, as well as their preferred method for receiving this information. Based on these results, we explore the potential for technology to bridge these information gaps and assist patients and caregivers in preventing errors in the inpatient environment. These contributions have meaningful implications for the design and development of systems to increase the engagement of both patients and caregivers in hospital safety.

**Methods**

To account for a broad range of patient and caregiver experiences, we created a survey with open-ended questions regarding undesirable events. We defined *undesirable event* within the survey as something that: was a small or big concern, was unpleasant or caused harm, and could have been avoided (e.g., not being informed that a scheduled procedure was cancelled, being given the wrong medication). Although these experiences might not typically be recognized as an error by clinicians, and might not match with a clinician’s perspective of what constitutes harm, we need to understand and recognize these events from a patient’s perspective. This study was approved by the authors’ Institutional Review Board.

**Site and Eligibility**

This study was conducted at Seattle Children’s Hospital, which manages over 300 beds, 5,000 hospital employees, and approximately 15,000 annual inpatient admissions. Patients were considered eligible if they were at least 7 years old, spent at least one night in the hospital during their current visit, and spoke English as their primary language. Caregivers were considered eligible if they were caring for a patient who had currently spent at least one night in the hospital, and if English was their primary language. Patients who were not well enough to give informed consent were not approached to complete the survey. We recruited participants from the acute care and intensive care units, and excluded those on the rehabilitation and inpatient psychiatric units due to the potential of our intervention to negatively affect safety, security, and patient care processes.

**Procedure**

We approached patients and caregivers after they spent at least one night in the hospital. A member of the research team described the study, defined the term undesirable event, and asked whether they had previously experienced such an event. If potential participants stated they had not experienced an undesirable event, they were deemed ineligible and thanked for their time. Those individuals who did experience an undesirable event were given further information about the survey instrument, and at that point either declined or consented to participate in this study. The consented participants were administered the survey instrument on an iPad, which was thoroughly wiped down with sanitizing cloths after each use to prevent the spread of infection within the hospital\(^21\). A research team member was available to answer questions or help the participant take the survey if necessary. Because we focus on patients’ perceptions, comparing the events that respondents mentioned in the survey with provider perspectives or official medical record reports was outside the scope of this study.

**Survey Instrument**

Our survey instrument was hosted by SurveyGizmo ([www.surveygizmo.com](http://www.surveygizmo.com)) and went through several iterations of content validity and testing to ensure that the questions were clear, thorough, and understandable for all eligible participants, including children as young as 7 years old. The survey questions related to the participant’s undesirable event were a combination of closed- and open-ended questions asking them: (1) what was the undesirable event that they experienced, (2) what they believed
caused this event, (3) what was the outcome of this event, (4) what they believe could have kept this event from happening, (5) what information the respondent thought they needed during this event, and (6) in what format they would prefer to receive this information. The participant could answer these questions about any previous undesirable event, not necessarily related to their current hospital stay. For example, caregivers could respond to the survey about an event in a different hospital that occurred several years ago with a family member other than their current pediatric patient. This flexibility helped to broaden the types of events we were interested in capturing.

**Data Analysis**

We analyzed the demographic and closed-ended survey question results using the analytical tools within SurveyGizmo. We conducted an inductive thematic content analysis on the responses to the four open-ended survey questions related to the respondent’s undesirable event. The research team met regularly during data collection to read new survey responses and revise the codebook based on emerging themes observed in the data set. Three members served as coders, assigning each survey response one or more codes that represented these themes. Data analysis concluded when themes in the codebook reached saturation and each completed open-ended survey response was assigned at least one code.

**Results**

Our qualitative analysis found four major types of undesirable events that patients and caregivers discussed in their survey responses: Mismanagement, Communication, Policy, and Lack of Care Coordination. For each of these categories, we provide explanations with quotes and anecdotes from our participants below, as well as summarize the outcomes associated with these events (Figure 1). We also examine what information participants thought would have been helpful in preventing or reducing the impact of these events (Figure 2), and how they prefer to receive this information. Patient and caregiver responses are designated with ‘P#’ and ‘C#’ identifiers, respectively. Because we also assigned identifiers to 10 blank and 13 incomplete survey responses that were excluded from our analysis, some identifier numbers are greater than the number of complete responses.

### Participants

Our research team approached 173 individuals in the hospital, of which 83 (47.98%) were considered ineligible because they had not experienced an undesirable event, 22 (12.72%) declined, and 68 (39.31%) completed the survey. Participant demographics are summarized in Table 1. Three of our participants expressed their desire to answer questions about more than one personal experience with undesirable events. Therefore, 68 individuals completed 71 survey responses describing 71 undesirable events, of which 16 (22.5%) identified being a patient, and 55 (77.5%) identified being a caregiver at the time of the undesirable event.

### Mismanagement

‘Mismanagement’ is a term we use to characterize undesirable events that consisted of clinicians’ errors in judgement or decision-making, and problems with the execution of a task or completion of a process. The strongest themes within this category dealt with medication and equipment errors.

### Medications

Patients and caregivers reported a variety of medication errors, including not receiving needed medications, receiving the wrong dosage, and receiving the wrong medication. One caregiver discussed her son not getting a medication for

---

**Table 1. Survey Response Rates and Demographics**

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey respondents*</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>Surveys completed</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Patients/caregivers approached</td>
<td>173</td>
<td></td>
</tr>
<tr>
<td>Patients/caregivers ineligible</td>
<td>83</td>
<td>48</td>
</tr>
<tr>
<td>Patients/caregivers declined</td>
<td>22</td>
<td>12.7</td>
</tr>
<tr>
<td>Caregiver responses</td>
<td>55</td>
<td>77.5</td>
</tr>
<tr>
<td>Patient responses</td>
<td>16</td>
<td>22.5</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 or over</td>
<td>62</td>
<td>87.3</td>
</tr>
<tr>
<td>14-17</td>
<td>7</td>
<td>9.9</td>
</tr>
<tr>
<td>7-13</td>
<td>2</td>
<td>2.8</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>13</td>
<td>18.3</td>
</tr>
<tr>
<td>Female</td>
<td>57</td>
<td>80.3</td>
</tr>
<tr>
<td>I prefer not to answer</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other - write in</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No answer</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Race/Ethnicity**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White or Caucasian</td>
<td>51</td>
<td>71.8</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>7</td>
<td>9.9</td>
</tr>
<tr>
<td>Black or African American</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Asian</td>
<td>7</td>
<td>9.9</td>
</tr>
<tr>
<td>Native American/American Indian</td>
<td>4</td>
<td>5.6</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other - write in</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>I prefer not to answer</td>
<td>7</td>
<td>9.9</td>
</tr>
<tr>
<td>No answer</td>
<td>1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

* 3 respondents completed multiple surveys about different undesirable events

** Participants could choose multiple categories
urinary retention for five days, resulting in the patient holding urine for over 12 hours at a time and the
need to put in a catheter (C53). Regarding medication administration, another caregiver wrote about an
event that happened to her daughter right before undergoing a major procedure: “She was given 5 times to
[sic] much morphine and she was hardly breathing” (C47). In other cases, patients received the wrong
medication intended for another patient. A pediatric patient recalled a mix-up while sharing a hospital room
with his sister: “Someone did not look at the label close enough” (P12).

Equipment
Participants also mentioned the misconfiguration or mishandling of medical devices. One NICU caregiver
noticed that the nurse did not use a board to secure her son’s IV. She later woke in the middle of the night
to find the patient’s IV detached and his hospital gown “soaked in blood.” She recalled: “I didn’t know
what had happened...but it looked pretty scary, and made me wonder when he had last been checked on by
staff” (C45). In another example, C63 woke suddenly for early morning rounds at 6:00 am. As the
examination began, this caregiver noticed the patient’s External Ventricular Drain was not clamped off, and
brought this oversight to the attention of the providers in the room. “At that point they realized they had the
incorrect patient. They were looking for the baby in the next bed. Her parents were not with her...I’m
thankful I came to my senses in time to interrupt them.” Because C63 noticed the problem and shared this
information with the providers, she stopped a treatment from being given to the wrong patient.

Outcomes
Many of the undesirable events were successfully resolved. For example, C53’s son eventually received the
necessary medication, while C47 noted her daughter’s mistake in morphine administration was reversed,
and the patient was able to wake up. Other situations, however, lead to a decline in the patient’s health,
which required additional care. For example, one patient was told she needed to remain flat after a
procedure, but was walked back to her hospital room instead. Severe headaches and vomiting developed as
a result. “My hospital stay was lengthened due to this and has caused lower back pain for several years”
(P94). As a result of these experiences, patients and caregivers reported feeling angry, scared, and worried.

Communication
Many undesirable events discussed by survey respondents involved communication problems that manifested
in different ways: between providers, in receiving information, and in providing information.

Between Providers
Participants identified conflicts and a lack of communication among their providers as a type of undesirable event. Regarding his
diagnosis, a pediatric patient said that “the doctors were disagreeing amongst themselves” and, in this patient’s opinion,
this disagreement caused them to overlook the data and miss “the real problem” (P30). In another example, a caregiver
described an event where disagreements between different teams of providers actually distracted them from the patient’s
increasingly urgent medical situation: “[my son] turned blue and was gasping for air. The nurse my son had that day took charge
and called the charged [sic] nurse for help. As we moved my son to the ICU floor, both team[s] were still arguing ” (C91).

Figure 1. Relationship between reported events and outcomes. We identified 143
unique relationships when coding the undesirable events. Mismanagement and
Communication events account for most of the self-reported events. While there
was more consistency between the different amount of outcomes, impacted care
plans and emotions were mentioned more frequently than other outcomes.
Not Receiving Information

Participants described difficulty in dealing with the lack of information. One parent mentioned, “A doctor I had never met before discussed worst case scenario treatment options in a very casual fashion...He was not a doctor in charge of making treatment decisions but more of a consult however didn’t identify himself that way” (C78). Because this caregiver did not receive more information about the provider, including why he was in the room, it influenced C78’s interpretation of treatment options that were discussed. Another caregiver remembered a time when the IV team was called to repair her son’s leaking line, but took over 2 hours to arrive. When asked about what she thought caused this event, C90 responded, “unwillingness of IV team to give updates despite repeated requests from [hospital department]”. This caregiver felt “caught off guard” because this delay also prevented the patient’s family from receiving critical information about his treatment side effects.

Challenges in Providing Information

Many caregivers felt that their voiced concerns about potential or ongoing problems were not acknowledged. For example, one NICU caregiver kept waking in the middle of the night due to the continuous alarm from the isolette. According to this caregiver, the nurse would come into the room and attempt to shut off the alarm every few minutes. “We were familiar with the isolette…and so could tell she had it on the wrong mode. We continued voicing our concern and the machine kept beeping...finally they had a NICU nurse come down and show her that she didn’t have the temperature probe on the baby’s body...” (C69). Although this caregiver expressed her concern to the nurse, she had to wait for a different provider to intervene and turn off the alarm. In another example, one adolescent patient described his unwillingness to give his providers information because he felt he was not receiving the care he needed: “the doctors [sic] didn’t give me anything to help the pain. It basically felt like they gave up on me and said suck it up...I soon didn’t even want to talk to the doctors [sic] at all because they were being horribly rude to me” (P18).

Outcomes

As a result of these communication errors, patients and caregivers experienced a range of negative emotions (e.g., frustration, stress, unhappiness, distrust, fear), health outcomes, delays in care or discharge, and need for additional care. Some respondents indicated that their undesirable events were successfully resolved. For example, C90’s son eventually had his IV line repaired. A few caregivers explicitly mentioned speaking up and requesting different providers to care for their child (C20, C65). Caregivers also expressed a desire to be more involved in the care process.

Policy

Patients and caregivers found themselves at odds with policies that seemed unreasonable or inconvenient. During a previous admission at a different hospital, P21 made a request for manual chest physiotherapy (CPT), a request that was not fulfilled due to “liability concerns”. The patient described, “I have had manual CPT at every other hospital I have ever been admitted to without any problem and found their policies to be ridiculous” (P21). One caregiver mentioned that hospital policy made it difficult to keep her family physically together: “After we had been asleep a while they came and told us hospital policy that our son cannot stay here but they were trying to approve it. So approve it! It’s been a long day and I don’t see how telling me that when we’re already in bed can help anyone” (C38).

Adverse effects of contact isolation protocols were also mentioned. A patient described her frustration with being placed under isolation when she did not believe she met its criteria. The patient’s siblings were not allowed to visit, and she was unable to receive the emotional support she wanted (P89). C62 was not allowed to store perishable food in the patient’s room according to contact isolation rules. Their room was far from the cafeteria or nutrition room, and an increased financial cost was associated with obtaining a room closer to food storage. As a result, both the caregiver and patient were unable to eat properly during their hospital stay.

Outcomes

Respondents cited negative health outcomes and negative emotions as a result of undesirable events related to policy. Because P21’s preferences were not accommodated, there was “barely any improvement in lung function, followed by another hospital stay for the same lung issues just a few months later.” Due to C62’s
lack of food access, she was unable to eat regularly, resulting in poorer general health. Negative emotions were also mentioned. P89 reported that contact isolation made her feel trapped in her room. “I felt as though [the doctors] didn’t understand how much it affected me as a patient.”

**Lack of Care Coordination**

The term ‘Lack of Care Coordination’ includes problems with scheduling, logistics, or coordination among providers, families, and patients. One participant described: “Care was not clustered and woke baby up repeatedly through the night, often just 30 minutes apart from one another” (C43). Two other caregivers mentioned missed opportunities to speak with providers about their child’s care (C57, C49). For C49, this prevented the patient’s parents from discussing their concerns about the patient’s care. “Further attempts were made by caregiver to state concerns. Hospitalist involved but damage was done.”

A lack of planning caused inconvenient delays in discharge for some patients. P39 stated, “I got discharged on Monday when my doctor said it could have been on Sunday” due to both a lack of coordination and communication. P16, an adolescent patient, experienced a discharge delay “for about a week” because both the patient and hospital staff had difficulty contacting the appropriate caregiver to sign the discharge paperwork. There was no plan in place to accommodate for this situation, and despite being healthy enough for discharge, the patient remained in the hospital until someone was available to provide a signature for her release.

**Outcomes**

In addition to discharge delays, C57’s daughter experienced a delay in care because he was not present for his child’s procedure. Negative emotions were also identified as outcomes of events related to lack of care coordination. C43’s child was “sad” and “cranky” the following day due to lack of sleep, while P16 struggled with unwanted feelings of being alone in the hospital. Respondents felt that proper planning and communication would have prevented these events from occurring. “Coordinate with each other...make sure you know what needs to be done in advance” (C43). C49 expressed a desire for “an agreed upon round time that met [the] schedule of all parties involved.”

**Information Needs and Preferences**

After describing their undesirable event, respondents were provided a list of 16 information needs, and asked if one or more of these items would have been helpful for them to know. The results of this question are shown in Figure 2. Differences in information needs were found across the types of events and between patients and caregivers. For example, patients who experienced mismanagement, communication, or policy errors were most interested in having information about their tests, procedures, and treatment plans. Patients who experienced care coordination problems wanted to know about discharge, while caregivers more interested in the timing of clinician visits. However, across all types of undesirable events that emerged from our analysis, caregivers most frequently indicated their desire to know ‘who to contact for help or questions’.

Respondents were then asked how they prefer to receive these information needs. They selected one or more modes of information delivery, from a list of 9 non-technology
based (e.g., paper handouts, written on the whiteboard, talking with doctors and nurses) and existing electronic (e.g., patient portal, text message, email) methods. Of the 80 total choices that our 16 patient respondents made, 68 (85.0%) fell into the non-technology based category, while 12 (15.0%) fell into the electronic category. Our 55 caregiver respondents made a total of 263 choices, of which 183 (69.6%) indicated non-electronic and 80 (30.4%) were electronic preferences for information delivery.

Discussion
Our participants provided informative data about the context and consequences of patient and caregiver experiences with undesirable events. Our results highlight the value of these perspectives in informing the design, development, and improvement of patient engagement technologies that can support error prevention within the hospital environment.

Undesirable Events from the Patient Perspective
Our findings extend previous definitions of undesirable events by accounting for patient and caregiver experiences. In addition to confirming previous concepts of undesirable events, participants also discussed mismanagement of situations regarding medications and medical equipment, disagreements with hospital policies, and challenges with communication and coordination. In related work, patients were usually asked to provide a binary response (yes or no) as to whether they encountered specific instances of clinician-defined undesirable events. These studies rarely allowed patients to provide further details. Our survey gave patients and caregivers the opportunity to share their knowledge in an unrestricted, open-ended way. This patient-focused approach revealed previously unidentified types of errors, and explored patient-perceived causal relationships between an undesirable event and its impact on the patient and caregiver (Figure 1). This knowledge permits future systems to better align with the needs of patients and caregivers.

Implications for Health IT and Inpatient Engagement Systems
Our results yield three major implications for inpatient information systems designed to increase patient engagement in the prevention of undesirable events. These systems should: (1) accommodate different information needs and preferences, (2) provide real-time information about the patient’s care, and (3) supplement face-to-face communication with clinicians.

Accommodate Different Information Needs and Preferences
Our findings indicate that information needs differed across types of events as well as between patients and caregivers. It is extremely important for patient-facing solutions to provide this range of information, due to current challenges in anticipating the nature of an undesirable event during a patient’s hospital stay. Future work in this area could further explore patterns of information needs associated with types of events, which would assist in the prediction or delivery of this tailored information to the patient and caregiver.

When asked about how they prefer to receive the information above, respondents favored more traditional communication options (e.g., paper handout, written on the patient’s whiteboard, talking with clinicians,) rather than modern solutions (e.g., patient portal, email, text message). This focus on non-technology based solutions could be explained by current hospital practices that tend to rely on such traditional solutions for distributing information, and the barriers patients and caregivers encounter when attempting to capture this information using technology. Despite our respondents’ strong preference for paper, handwritten, and verbal communication, recent work involving pediatric inpatient portals found that 89% of parents thought the use of such a portal helped reduce errors in the patient’s care. More research is needed to understand how the availability and enhancement of similar tools can impact patient and caregiver preferences for information delivery.

Provide Real-Time Information
Patient-facing information systems for the inpatient environment need to provide patients and caregivers with real-time information about the patient’s care. One respondent, a caregiver, described a situation where she did not receive updates about changes made to the patient’s medication plan: “[there were] too many people involved and too many changes made and parents not aware of them” (C15). In this case, not receiving updates about changes to the patient’s medication was the primary reason for the caregiver’s undesirable event. Because undesirable events can arise at any point during a hospital stay, we strongly recommend that systems provide patients and their caregivers with information about their care without any
built-in time delay, providing capture and access functionality that supports retrieval and review at a patient or caregiver’s convenience. These tools have the potential to share information with other family members and caregivers who make critical decisions about the patient’s care. Also, providing additional information (e.g., explaining what a lab result means) allows patients and caregivers to better understand treatments and ultimately strengthen their ability to act as their own advocate.

Supplement Face-to-Face Communication
Inpatient systems also have the potential to offer patients and caregivers alternative communication routes when face-to-face communication fails. Many respondents described different ways in which these failures occurred. For example, one caregiver mentioned a conflict with a nutritionist regarding her son’s diet. The nutritionist “did not listen to what I had to say”, and ultimately the caregiver “got irritated and took matter[s] into my own hands” (C91). In this case, when in-person communication broke down, an electronic system could have allowed C91 to non-confrontationally request support or opinions from another clinician. Another caregiver discussed a time where face-to-face communication with her son’s providers actually caused increased worry and concern: “The resident came in [and] was asking questions that were causing me alarm. It was making me worried about my son and if there was something going on that they weren’t telling me…” (C51). In situations like this, when there is increased potential for miscommunication, or when there is an evident lack of information, inpatient tools can serve as a mechanism for patients or caregivers to request further details and ask follow-up questions when needed.

Limitations and Future Work
Individuals who were not eligible, not well enough to provide consent or answer survey questions, or declined participation were not given the survey. Because of this constraint, other types of undesirable events could have been missed. Three participants responded to our survey more than once, because they expressed a desire to answer questions about multiple undesirable events. Although this introduced a demographic redundancy in our data set, we independently tracked the number of unique individuals we approached during this study, and have reported the differences in these values. A selection bias could influence our results, as respondents were not offered compensation for their participation, and those who were frustrated with their situation might have had increased motivation to participate. Our sample size and single hospital location also limits the generalizability of our results. Pediatric patients were less likely to complete our survey due to illness or ineligibility. Despite their reduced involvement in our study, the diverse experiences of this population indicates an area of further exploration. A better understanding of the perceived usefulness of existing inpatient technologies, and their current or potential role in preventing medical errors is also needed.

Conclusion
In this paper, we presented the results of a survey asking pediatric inpatients and their caregivers about their prior experiences with undesirable events. We categorized the perceived types of events and illustrated their relationship with outcomes that patients and caregivers described. Our findings demonstrate that patients and caregivers have a broad definition of undesirable events, have many unmet information needs that they feel might help prevent such events, and have specific preferences for receiving this information. Based on these results, we made recommendations for future iterations of patient engagement technology, which has the potential to increase support for patients and caregivers in becoming active participants in their healthcare, and help reduce the tragic toll of medical errors.

Acknowledgements
We would like to thank each of our participants for taking time during their hospital stay to take part in this study, the Patients as Safeguards research team, Kathryn Nickel, and members of the Outcomes Assessment Program at Seattle Children’s Hospital (Stephanie Hillman, Emily Dehmer, Jessica Kowalchuck, Kristine Lee, Neil Panlasigui, and Breanna Dolan) for their partnership and support. This research study was primarily funded by the Agency for Healthcare Research and Quality (AHRQ) grant #1R01HS022894, with additional support from the National Library of Medicine Biomedical and Health Informatics Training Grant #T15LM007442.
References

4. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ. 2016 May 3; 353:i2139
Topological-Pattern-Based Recommendation of UMLS Concepts for National Cancer Institute Thesaurus

Zhe He, PhD1,2, Yan Chen, PhD3, Sherri de Coronado, MS, MBA4, Katrina Piskorski, MD5, James Geller, PhD6

1School of Information, Florida State University, Tallahassee, FL; 2Institute for Successful Longevity, Florida State University, Tallahassee, FL; 3Department of Computer Information Systems, Borough of Manhattan Community College, City University of New York, New York, NY; 4National Cancer Institute, Rockville, MD; 5Weill Cornell Medicine, New York, NY; 6Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

Abstract

The National Cancer Institute Thesaurus (NCIt) is a reference terminology used to support clinical, translational and basic research as well as administrative activities. As medical knowledge evolves, concepts that might be missing from a particular needed subdomain are regularly added to the NCIt. However, terminology development is known to be labor-intensive and error-prone. Therefore, cost-effective semi-automated methods for identifying potentially missing concepts would be useful to terminology curators. Previously, we have developed a structural methodology leveraging the native term mappings of the Unified Medical Language System to identify potential concepts in several of its source vocabularies to enrich the SNOMED CT. In this paper, we tested an analogous method for NCIt. Concepts from eight UMLS source terminologies were identified as possibilities to enrich NCIt’s conceptual content.

Introduction

Biomedical ontologies and controlled terminologies provide a solid foundation in a variety of healthcare information systems [1, 2]. They have been widely used for encoding diagnoses, laboratory tests, and problem lists in Electronic Health Records [3] as well as in administrative documents such as in billing statements [4]. Moreover, with medical concepts linked by taxonomic and semantic (lateral) relationships, they also play an important role in knowledge management, data integration, and decision support [1]. Complicated and challenging natural language processing tasks also benefit from well-curated domain ontologies and controlled terminologies.

The National Cancer Institute thesaurus (NCIt), accessible through a browser at https://ncit.nci.nih.gov/ncitbrowser/ is a reference terminology developed and maintained by the National Cancer Institute (NCI). Currently, it contains over 100,000 concepts that are hierarchically organized in 19 distinct domains related to cancer research, e.g., neoplastic diseases, molecular abnormalities, and genes. It is a central reference terminology of NCI’s Enterprise Vocabulary Services (EVS) [5]. As medical terminology is constantly evolving, with new concepts in healthcare entering usage, a controlled terminology needs to keep improving its conceptual content to encode these new concepts as they are needed by users. In Cimino’s “desiderata” for controlled medical vocabularies [6], domain completeness is listed as the most desirable property. To improve the conceptual content of NCIt, NCI EVS exploits both its internal quality assurance (QA) mechanisms as well as external participation in the development and QA process of NCIt. A contributor outside of the NCI can suggest new needed terms or larger sets of concepts for NCIt that will be subsequently reviewed by EVS, validated and developed in conformance with NCIt content development and editing guidelines. The monthly update cycle of NCIt also ensures the timely incorporation of new terms.

In previous research, we have introduced a structural methodology to recommend new concepts from a UMLS source vocabulary for inclusion in another source vocabulary where they are “missing” [7, 8]. This algorithmic structural method explores vertical density differences between pairs of terminologies in the Unified Medical Language System (UMLS). The method consists of recognizing trapezoids in the “parent of” relationship structures of terminologies, that is, recognizing cases where concept pairs are present in two terminologies but each terminology offers different intermediate concepts along their "parent of" paths of relationships. It leverages the native term mappings of the UMLS to identify topological patterns that are indicative of a possible import of a concept from one terminology into another terminology. Examples of topological patterns will be shown in the Background and Methods sections.
A variety of vertical topological patterns (‘trapezoids’ of different sizes) were used in previous research to identify a list of candidate concepts for import into SNOMED CT. Human domain experts confirmed the validity of this structural method for enriching the conceptual content of SNOMED CT [7, 8]. In this paper, we apply this topological pattern-based method to recommend new concepts for inclusion in the NCIt. Just as in our previous research, it was hypothesized that a structural difference between two terminologies might suggest a different course of action besides an import, e.g., it might lead to uncovering an error in one of the two source terminologies. Different possibilities will be exhaustively enumerated in this paper.

**Background**

Quality assurance (QA) of the NCIt has been conducted by both NCI and external researchers [5]. Min et al. constructed an area taxonomy and a partial-area taxonomy for the NCIt that highlighted potential errors for manual review by a human expert [9]. Cohen et al. performed an automated comparative audit of the gene hierarchy of NCIt using the Entrez Gene database of the National Center for Biotechnology Information [10]. More recently, Semantic Web technologies have been leveraged to audit the NCIt. Mougin and Bodenreider stored the NCIt concepts in an RDF triple store to assess the consistency of the hierarchical and associative relationships among them [11]. Jiang et al. evaluated the data quality of cancer study common data elements by integrating the NCI Cancer Data Standards Repository, NCIt, and the UMLS Semantic Network with the use of a variety of tools of the Semantic Web [12].

The UMLS is a medical terminological system. Its Metathesaurus integrates over 12 million terms from more than 190 source vocabularies into about 3.25 million concepts, such that terms with the same meaning are assigned the same Concept Unique Identifier (CUI). Due to its large scale, source integration and term mapping are challenging tasks. To aid UMLS source integration, Huang et al. developed an extrinsic method that uses WordNet synonym substitution and showed promising results [13]. Moreover, the syntactic patterns and the semantics of the UMLS have been exploited, supporting ontology alignment for OBO Foundry ontologies [14].

The use of topological patterns was introduced in our previous research [7, 8] and will now be explicated based on “k:1” trapezoids. Figure 1 shows the topological pattern of a k:1 trapezoid with k=2. The instances of the concept A have the same UMLS CUI (Concept Unique Identifier) in both terminologies, which means that the UMLS curators regarded them as the same concept. The same is true for the concept B. The identity of the “two” concepts A in the two terminologies is hinted at by the double line connecting them that reminds of the = symbol. However, Terminology 1 has an additional concept X located on a path from B to A. In our previous work, SNOMED CT was the terminology of interest. Encouraged by the success with SNOMED CT, we are taking steps towards establishing the generality of the method by using a different target terminology, namely the NCIt. Therefore, we will formulate our explanation in terms of the NCIt.

Looking at Figure 1, one can argue that in the limited scope of paths from B to A, Terminology 1 contains more information than the NCIt, by including the additional concept X. The concept X is not just missing on the path from B to A in NCIt, but we ascertain that X does not appear anywhere in NCIt before we consider importing it. The topological pattern illustrated in Figure 1 is referred to as 2:1 trapezoid, because there are two parent-child links on the Terminology 1 side and one parent-child link on the NCIt side with links of equal length connecting A with A and B with B.

Figure 1. Abstraction of a 2:1 trapezoid between Terminology 1 and Terminology 2 (NCIt). Should X be imported into the NCIt?
Methods

In this work, we use the term “candidate terminology” to refer to any selected UMLS source terminology that could potentially contribute concepts to NCIt (version 2014_03E). Candidate terminologies were determined as follows. We first identified all English source terminologies with “PAR” (Parent) relationships annotated with “INVERSE_IS_A” labels that overlap the content of NCIt. The rationale for excluding “RB” (Broader) relationships in this study is based on the fact that “PAR” represents an explicitly defined parent-child relationship in the source, whereas “RB” represents an implicit one, inserted by the UMLS editorial team [15]. In the UMLS 2015AA release, we identified CPM, CPT, FMA, GO, MEDCIN, SNOMEDCT_VET, ATC, and UMD as the candidate terminologies. Table 1 lists the versions and full names of these terminologies. In previous research [7, 8], we have used this method to identify potential concepts in NCIt to enrich SNOMED CT, and vice versa. Thus, in this work, we excluded SNOMED CT from the list of candidate terminologies even though it also has “PAR” relationships annotated with “INVERSE_IS_A” labels.

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Version</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>December 17, 2014</td>
<td>MEDCIN</td>
</tr>
<tr>
<td>Foundational Model of Anatomy Ontology</td>
<td>3.1</td>
<td>FMA</td>
</tr>
<tr>
<td>Gene ontology</td>
<td>May 19, 2014</td>
<td>GO</td>
</tr>
<tr>
<td>Medical Entity Dictionary</td>
<td>2003</td>
<td>CPM</td>
</tr>
<tr>
<td>Universal Medical Device Nomenclature System</td>
<td>2015</td>
<td>UMD</td>
</tr>
<tr>
<td>Anatomical Therapeutic Chemical classification system</td>
<td>March 2, 2015</td>
<td>ATC</td>
</tr>
<tr>
<td>SNOMED CT Veteran Extension</td>
<td>October 20, 2014</td>
<td>SNOMEDCT_VET</td>
</tr>
</tbody>
</table>

The topological patterns considered in this paper are $k:1$ ($k = 2, 3, 4, \ldots$) trapezoids, 3:2 and 2:3 trapezoids as well as 2-rectangles. The definitions of these topological patterns are given below.

**K:1 Trapezoids**

In this study, Terminology 1 is always one of the candidate terminologies listed in Table 1. Discovery of a $2:1$ trapezoid suggests that the concept $X$ could be imported into the NCIt. It has been our experience that the owners of several different terminologies have expressed reasons why not to include such new concepts. The main justifications were that “the concepts describe intermediate, non-coding terms or alternative classifications that would not add substantial value” and “we cannot include everything, if we start thinking about what we could include, it never ends.” Thus the final decision whether “X” should be included in NCIt always has to be made by its curators. An alternative classification recognizes the fact that not every concept in one terminology is valid for another, because the two terminologies might have different purposes, or because the concept is part of a different classification schema in the two ontologies. A technical explanation of alternative classification is given below.

We implemented a program that generates all possible $k:1$ ($k = 2, 3, 4, \ldots$) trapezoids for all terminologies from Table 1 with the following termination condition for the size of $k$. For a given terminology $T$ from Table 1, if no $k_0:1$ trapezoid is found, then the algorithm checks whether a $(k_0+1):1$ trapezoid exists. If the answer is “no,” the algorithm terminates for $T$ and continues to the next terminology until all terminologies have been processed. We note that determining the largest value of $k$ for which $k:1$ trapezoids exist is of practical and “academic” interest. For a $k:1$ trapezoid, the $k – 1$ concepts on the path from the concept $B$ to the concept $A$ are then proposed by the algorithm as possible imports into the NCIt to the human expert. The algorithms were published previously [7,8].

**Experiment 1:** For a sample of 2:1 trapezoids we gave our oncology expert (KP) the following choices:

1) The intermediate concept in Terminology 1 (e.g., Concept X in Figure 1) should be imported into NCIt.
2) The intermediate concept in Terminology 1 should not be imported into NCIt because it is not relevant to cancer.
3) The concept structure is incorrect to begin with (e.g., Concept X should not be a child of Concept A).
4) Other (please fill in).

The distribution of concepts between these four choices will be shown in the Results section.
**M:N Trapezoids and M-Rectangles**

*M:N* trapezoids are a generalization of *k:1* trapezoids. Whenever *M = N*, it is geometrically more appropriate to refer to an *M*-rectangle. We formally define 2-rectangles and *M:N* trapezoids as follows:

**Definition 1**: The concepts *A*, *B*, and *X* (from Terminology 1) and *A*, *B*, and *Y* (of NCIt) form a 2-rectangle if and only if:

- The concepts *X* and *Y* have identical parents in Terminology 1 and in the NCIt (in this case *A*).
- The concepts *X* and *Y* have identical children in Terminology 1 and in the NCIt (in this case *B*).
- The concept *X* does not appear anywhere in the NCIt.
- The concept *Y* does not appear anywhere in Terminology 1.
- There is no synonymy relationship and no hierarchical relationship between *X* and *Y* known (in the UMLS).

**Definition 2**: The concepts *A*, *B*, and *X* (from Terminology 1) and *A*, *B*, and *Y* (of NCIt) form an *M:N* trapezoid if and only if:

- The concepts *X* and *Y* have identical parents in Terminology 1 and in the NCIt (in this case *A*).
- The concepts *X* and *Y* have identical children in Terminology 1 and in the NCIt (in this case *B*).
- The concepts *X* do not appear anywhere in the NCIt (*i.e.*, *X* = 1..*M*-1).
- The concepts *Y* do not appear anywhere in Terminology 1 (*i.e.*, *j* = 1..*N*-1).
- There is no synonymy relationship and no hierarchical relationship between the *X* and *Y* known (in the UMLS).
- Every concept *X* is a child of the concept *X* and as parent of *B*. Every concept *Y* is a child of the concept *Y*.

Figure 2 shows an abstract layout of a 2-rectangle to elucidate the above Definition 1.

![Figure 2](image)

*Figure 2*. (a) An abstract layout of a 2-rectangle topological pattern. The double-headed arrow indicates the case that *X* is a synonym of *Y*. The no-fill, one-headed arrow indicates that *X* could be imported into the NCIt as child of *A* and as parent of *Y*. The solid-fill one-headed arrow indicates that *X* could be imported into the NCIt as child of *Y* and as parent of *B*. This scenario is shown in (b). Only one of these three cases is possible. In addition it is possible that there is an error, *e.g.*, *Y* is not really a child of *A*.

In a 2-rectangle there are six possible cases for how *X* and *Y* may relate to each other.

1) It holds that NCIt could contain the path of concepts connected by links *B* → *Y* → *X* → *A*. This import case is hinted at by the no-fill, one-headed arrow in Figure 2(a).
2) It holds that NCIt could contain the path of concepts *B* → *X* → *Y* → *A*. This import case is hinted at by the solid-fill, one-headed arrow in Figure 2(a). The resulting path of this import is shown in Figure 2(b).
3) Concept *X* is a real world synonym of concept *Y*, which was previously not recognized by the UMLS editors. In other words, it holds that *X* = *Y*. More precisely, it holds that CUI(*X*) = CUI(*Y*), using the symbol for identity.
4) There might be a structural error in Terminology 1, *e.g.*, *X* is not really a child of *A*.
5) There might be a structural error in NCIt.
6) The concepts *X* and *Y* are *alternative classifications*. This is the most difficult case from a conceptual standpoint. It indicates two different ways of how to conceptualize a domain that are both valid but not immediately compatible. This is best understood with an artificial example. Assume that *A* = *extremities* and *B* = *left leg*. Then *X* may validly be *upper extremities* while *Y* may be *left extremities*. These are two different, valid ways how to organize the immediate subclasses of *A*, however, a deeper analysis is required to make the two approaches compatible. This case is of limited interest in the current paper.
We are primarily interested in cases 1) and 2), while case 3) may also provide an enrichment of NCIt. We note, parenthetically, that Y might also be imported into Terminology 1 (Figure 2).

**Experiment 2a:** A sample of 2-rectangles was reviewed by YC. YC was given the six choices above.

**Experiment 2b:** SDC reviewed the results of Experiment 2a and commented on them.

We will now advance from 2-rectangles to 3:2 trapezoids. Again, Concept A and Concept B are identical in both Terminology 1 and NCIt. There are two intermediate concepts (Z and Y) on a path from B to A in Terminology 1, while there is one intermediate concept (X) between Concept B and Concept A in NCIt.

**Experiment 3:** A sample of 3:2 trapezoids was reviewed by YC. YC was given eight choices.

1) X and Y are alternative classifications of A. This is analogous to alternative classifications for 2-rectangles.
2) Y and Z are both imported into NCIt and it holds that $B \rightarrow Z \rightarrow Y \rightarrow X \rightarrow A$.
3) Y and Z are both imported into NCIt and it holds that $B \rightarrow Z \rightarrow X \rightarrow Y \rightarrow A$.
4) Y and Z are both imported into NCIt and it holds that $B \rightarrow X \rightarrow Z \rightarrow Y \rightarrow A$.
5) $Y = X$ (Z could also be a child of X)
6) $Z = X$ (Y could also be a parent of X)
7) There might be a structural error in Terminology 1, e.g., Z is a synonym of Y.
8) There might be a structural error in NCIt, e.g., X is not a child of A at all.

We limited the burden for the reviewer by excluding the consideration whether the concepts Y and Z are desirable for import into a cancer terminology, otherwise additional cases would arise, such as:

9) Only Y is a desirable import, and it holds that $B \rightarrow Y \rightarrow X \rightarrow A$.
10) Only Y is a desirable import, and it holds that $B \rightarrow X \rightarrow Y \rightarrow A$.
11) Only Z is a desirable import, and it holds that $B \rightarrow Z \rightarrow X \rightarrow A$.
12) Only Z is a desirable import, and it holds that $B \rightarrow X \rightarrow Z \rightarrow A$.
13) Neither Y nor Z is desirable to be imported into the NCIt cancer terminology.

For a 2:3 trapezoid, analogous cases can be defined, which was done in this research, even though the primary interest is in importing into NCIt. A 2:3 trapezoid would indicate an opportunity of exporting concepts from NCIt.

**Experiment 4:** A sample of 2:3 trapezoids was reviewed by YC. YC was given eight choices as in Experiment 3.

The UMLS contains cycles of hierarchical relationships [15]. Furthermore, multiple parents may lead to overlapping trapezoids which could lead to counting the same intermediate concept multiple times. We eliminated cycles (by detecting the repetition of a CUI along a PAR path) as well as duplicate intermediate concepts in the results.

**Results**

Table 2 shows the numbers of potential import concepts identified in $k$:1 trapezoids. When calculating the numbers in Column 3, we made sure that these are unique concepts (and therefore there might be fewer concepts than trapezoids). Column 4 lists the total numbers of $k$:1 trapezoids found by the algorithm. Among the eight candidate terminologies, MEDCIN could contribute the largest number of concepts to the NCIt, followed by FMA and GO.

<table>
<thead>
<tr>
<th>Candidate terminology</th>
<th>Size of candidate terminology (# of CUIs)</th>
<th>Potential concepts in candidate terminology</th>
<th>Number of $k$:1 trapezoids</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>318,647</td>
<td>288</td>
<td>413</td>
</tr>
<tr>
<td>FMA</td>
<td>82,043</td>
<td>156</td>
<td>228</td>
</tr>
<tr>
<td>GO</td>
<td>57,226</td>
<td>154</td>
<td>179</td>
</tr>
<tr>
<td>CPT</td>
<td>38,975</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>SNOMEDCT_VET</td>
<td>36,032</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>CPM</td>
<td>3,077</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>UMD</td>
<td>21,794</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ATC</td>
<td>5,204</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3 shows the number of observed $k$:1 trapezoids ordered by increasing values of $k$. The table shows that $k$:1 trapezoids were found with $k$ up to 9. Both the number of trapezoids and the number of potential import concepts decrease with an increasing value of $k$, which is consistent with our previous work for enriching SNOMED CT [8]. For 2:1 trapezoids, we chose a random sample of 30 trapezoids.
Table 3. Number of $k:1$ trapezoids of each kind and corresponding number of potential import concepts.

<table>
<thead>
<tr>
<th>Kinds of trapezoids</th>
<th>Number of trapezoids</th>
<th>Potential concepts in candidate terminologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:1</td>
<td>520</td>
<td>314</td>
</tr>
<tr>
<td>3:1</td>
<td>153</td>
<td>160</td>
</tr>
<tr>
<td>4:1</td>
<td>75</td>
<td>106</td>
</tr>
<tr>
<td>5:1</td>
<td>63</td>
<td>102</td>
</tr>
<tr>
<td>6:1</td>
<td>26</td>
<td>50</td>
</tr>
<tr>
<td>7:1</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>8:1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>9:1</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>10:1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11:1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Of the 30 randomly chosen 2:1 trapezoids reviewed by the domain expert (KP), the intermediate concepts in 22 (73.3%) trapezoids were recommended for potential import into NCIt, whereas the intermediate concepts in 8 (26.7%) trapezoids were not recommended for import into NCIt. No trapezoids were assigned the choices 3) and 4) in Experiment 1. Among the 22 2:1 trapezoids, KP recommended that the intermediate concepts in three of them should be imported into NCIt with a variation of their original term. For example, “disorders of peripheral nerve, neuromuscular junction and muscle” (C2102996) was recommended to be imported between “nervous system disorder” (C0027765) and “peripheral neuropathy” (C0031117), but with a simplified string, which SDC changed to “peripheral nervous system disorder.” Table 4 shows three example trapezoids in which intermediate concepts were recommended for potential import into NCIt.

Among the eight 2:1 trapezoids from which the intermediate concept was not recommended for import into NCIt, six are from MEDCINE, one is from CPM, and the last one is from FMA. Four out of eight intermediate concepts are assigned the semantic type “Pharmacologic Substance.” One example where import was not recommended consists of A = “Muscle relaxants,” B = “Baclofen” and X = “Skeletal muscle relaxants.” Investigating the definition of Muscle relaxants as it is given in the NCIt, it turns out that it is defined as “Any agent that relaxes skeletal muscles and reduces muscle contraction,” thus its semantics is equal to the semantics of X itself. Thus, importing X would be redundant from the point of view of the NCIt.

Table 4. Example $k:1$ trapezoids in which the intermediate concept in the candidate terminology was recommended.

<table>
<thead>
<tr>
<th>Candidate terminology</th>
<th>NCIt</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:1</td>
<td></td>
</tr>
<tr>
<td>MEDCIN</td>
<td></td>
</tr>
<tr>
<td>Lymphoid Tissue (C0024296)</td>
<td>Lymphoid Tissue (C0024296)</td>
</tr>
<tr>
<td>Epithelium-associated lymphoid tissue (C1179414)</td>
<td>mucosa-associated lymphoid tissue (C0599921)</td>
</tr>
<tr>
<td>mucosa-associated lymphoid tissue (C0599921)</td>
<td></td>
</tr>
<tr>
<td>2:1</td>
<td></td>
</tr>
<tr>
<td>GO</td>
<td></td>
</tr>
<tr>
<td>Cell Cycle Checkpoints (C1155874)</td>
<td>Cell Cycle Checkpoints (C1155874)</td>
</tr>
<tr>
<td>mitotic cell cycle checkpoint (C2263179)</td>
<td></td>
</tr>
<tr>
<td>Mitotic Spindle Checkpoints (C1155750)</td>
<td>Mitotic Spindle Checkpoints (C1155750)</td>
</tr>
<tr>
<td>2:1</td>
<td></td>
</tr>
<tr>
<td>FMA</td>
<td></td>
</tr>
<tr>
<td>Epithelial Cells (C0014597)</td>
<td>Epithelial Cells (C0014597)</td>
</tr>
<tr>
<td>Endo-epithelial cell (C1181294)</td>
<td></td>
</tr>
<tr>
<td>Thymic epithelial cell (C0229951)</td>
<td>Thymic epithelial cell (C0229951)</td>
</tr>
</tbody>
</table>

Table 5 shows the numbers of 2-rectangles, 2:3 trapezoids and 3:2 trapezoids found by our algorithm. In order to analyze the relationships of intermediate concepts in the trapezoids, random samples of two times 30 trapezoids and 30 rectangles were chosen from MEDCIN. That yielded a total of 90 topological patterns for analysis. For all the other candidate terminologies, the domain expert (YC) reviewed all of the trapezoids that the algorithm identified. In total, 100 2-rectangles, 69 2:3 trapezoids, and 71 3:2 trapezoids were reviewed. Subsequently, SDC of the National
Cancer Institute further reviewed the marked up sample of 100 2-rectangles. We measured the inter-rater agreement between YC and SDC using Cohen’s Kappa. The observed Kappa is 0.6175, indicating substantial agreement [16].

Table 6 shows the human review results of the sample of 2-rectangles. All cases except “Error in NCIt” were observed. The results show that 55% of the rectangles in the sample are alternative classifications. Another 20% + 17% = 37% fall into two categories where the intermediate concept in the candidate terminologies could be imported into the NCIt as a parent or child of its intermediate concept. For comparison, in our previous study on SNOMED CT [7], 23.6% of the 2-rectangles in the sample fell into these two categories. Thus, the NCIt is a better target terminology for concept import than SNOMED CT, which may be due to the fact that SNOMED CT is more comprehensive than the NCIt. The percentage of cases of synonymy is 6.0% in this study, which is lower than the 14.5% in our previous study on SNOMED CT [7].

![Figure 3](image)

**Figure 3.** An example 2-rectangle in which the intermediate concept in MEDCIN could be a child of the intermediate concept in NCIt. Neutropenia, a condition in which there is a lower-than-normal number of neutrophils (neutrophilic white blood cell), is a kind of non-neoplastic hematologic and lymphocytic disorder.

![Hematological Disease (C0018939)](image)  ![Hematological Disease (C0018939)](image)  ![Non-Neoplastic Hematologic and Lymphocytic Disorder (C1518374)](image)  ![Cyclic neutropenia (C0221023)](image)  ![Cyclic neutropenia (C0221023)](image)  ![MEDCIN](image)  ![NCIt](image)
Figure 4 shows another example 2-rectangle where the intermediate concept in FMA “White matter of telencephalon” (C2327688) was deemed to be an alternative classification of the intermediate concept in the NCIt “Brain White Matter” (C1706995). In another 2-rectangle, two intermediate concepts “Implantable prosthesis” (C0021102) and “Implants” (C0012202) were deemed to be synonymous by the human evaluator.

Table 7 shows four examples of 2:3 and 3:2 trapezoids. In the first example, the intermediate concept in MEDCIN “Ovarian Carcinoma” (C0029925) was deemed to be a synonym of “Epithelial ovarian cancer” (C0677886). In the second example, two intermediate concepts “oncologic disorders” (C3853943) and “cancer-related problem/condition” (C0280950) were deemed to be alternative classifications. The reason that these are not synonyms is that the hierarchy of oncologic disorders in MEDCIN is used to represent the cancer diagnoses, which is not the same as the meaning of cancer-related conditions (the NCIt term). In the third example, the second intermediate concept “inherited genetic conditions” (C3648695) was deemed to be synonymous with “Hereditary Diseases” (C0019247). The fourth example is another case that is best understood as an alternative classification. Subclavian artery and Axillary Artery are Systemic arteries, but NCIt does not classify arteries as systemic.

Table 8 and Table 9 show the review results of the 2:3 and 3:2 trapezoids. No errors in the sample were observed. The results show that “alternative classification” is the most prevalent case, followed by “possible import as parent or child,” and then “synonyms.” These findings are also consistent with our previous study on SNOMED CT [8].
Table 8. Human review results of 2:3 trapezoids.

<table>
<thead>
<tr>
<th>Candidate terminology</th>
<th>Sample size</th>
<th>Alter. classification</th>
<th>Z \rightarrow Y \rightarrow X</th>
<th>Z \rightarrow X \rightarrow Y</th>
<th>X \rightarrow Z \rightarrow Y</th>
<th>X is a synonym of Y</th>
<th>X is a synonym of Z</th>
<th>Error in Term. 1 or NCIt</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>30</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>GO</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATC</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CPM</td>
<td>11</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CPT</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SNOMED CT VET</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FMA</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>UMD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>23</td>
<td>15</td>
<td>0</td>
<td>18</td>
<td>5</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Percentage</td>
<td>100%</td>
<td>33.3%</td>
<td>21.7%</td>
<td>0%</td>
<td>26.1%</td>
<td>7.2%</td>
<td>11.6%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 9. Human review results of 3:2 trapezoids.

<table>
<thead>
<tr>
<th>Candidate terminology</th>
<th>Sample size</th>
<th>Alter. classification</th>
<th>Y \rightarrow X \rightarrow Z</th>
<th>Y \rightarrow Z \rightarrow X</th>
<th>Z \rightarrow Y \rightarrow X</th>
<th>Z is a synonym of X</th>
<th>Z is a synonym of Y</th>
<th>Error in Term. 1 or NCIt</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDCIN</td>
<td>30</td>
<td>17</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>GO</td>
<td>25</td>
<td>16</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ATC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CPM</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CPT</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SNOMED CT VET</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FMA</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UMD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td>38</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Percentage</td>
<td>100%</td>
<td>53.5%</td>
<td>7.0%</td>
<td>11.3%</td>
<td>5.6%</td>
<td>9.9%</td>
<td>12.7%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Discussion and Conclusions

In this work, we applied a topological-pattern-based method to recommend concepts from eight UMLS source terminologies that could potentially enrich the NCIt’s conceptual content. A variety of topological patterns between pairs of UMLS source terminologies were identified by our algorithm with potential import concepts for expert review. Three domain experts reviewed the samples and determined whether the potential concepts in the samples should be imported and suggested how they should be inserted into existing structures of the NCIt. The results demonstrated the effectiveness of the topological-pattern-based method for enriching the NCIt. For 2-rectangles, 2:3 trapezoids, and 3:2 trapezoids, the prevalence of possible relationships between intermediate concepts in pairs of terminologies was consistent with our previous studies on SNOMED CT [8]. The most prevalent cases were alternative classification, followed by various forms of import.

A few limitations need to be noted. In this study, only “PAR” links with “INVERSE_IS_A” annotations were used. We did not look for rectangles larger than 2-rectangles, and we did not attempt to identify 4:2, 4:3, etc. trapezoids. As shown above, even a 3:2 trapezoid leads to a remarkable number of possibilities as to how intermediate concepts could relate to each other. For $M:N$ trapezoids with larger values of $M$ and $N$ there will be a combinatorial explosion of possible cases; that will make it difficult for a human expert to consider all of them even for a small number of trapezoids. Human review is important, however. In [17], we have conducted a preliminary analysis of the difficulty of importing pattern-based concepts into the NCIt. The contexts and definitions of potential new concepts originating from a source need to be evaluated to validate the intended meaning. In a few cases, SDC “overrode” the decisions of YC. We note that the viewpoints of an outside auditor and a curator of a terminology are necessarily different. A curator has a deeper knowledge and better understanding of how individual representational decisions
were reached. We also note that not all the potential concepts identified would be imported, because of the different uses and alternative schemas of different terminologies. However, our method provides a way to look at possibly needed areas for improving domain coverage. In future, it could play a role in regular QA of terminologies at NCI.

In future work, we will investigate the use of the NCI Metathesaurus instead of the UMLS Metathesaurus. It is a subset of UMLS with other sources added, and is well curated in areas of interest to NCI [18]. Moreover, NCIt is more frequently updated in the NCI Metathesaurus than in the UMLS. Another useful source for future work is the NCIt GO mapping of the NCI [19]. We plan to extensively use feedback of the NCI EVS. After getting feedback and arguments for inclusion and exclusion of our suggested concepts, we will refine our methods to produce more accurate recommended concepts accordingly. We also plan to investigate horizontal density differences with varying numbers of sibling concepts between pairs of terminologies to identify more missing concepts for import. Lastly we will work on 3:1, 4:1, 5:1 and 6:1, and 4:2 samples.

Acknowledgments

Research reported in this publication was partially supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Adolescent and Caregiver use of a Tethered Personal Health Record System

Matthew K. Hong, MS¹, Lauren Wilcox, PhD¹, Clayton Feustel, MS¹, Karen Wasileski-Masker, MD, MSc², Thomas A. Olson, MD², Stephen F. Simoneaux, MD²

¹Georgia Institute of Technology, Atlanta, GA
²Children’s Healthcare of Atlanta, Atlanta, GA

Abstract

Supporting adolescent patient engagement in care is an important yet underexplored topic in consumer health informatics. Personal Health Records (PHRs) show potential, but designing PHR systems to accommodate both emerging adults and their parents is challenging. We conducted a mixed-methods study with teenage adolescent patients (ages 13-17) with cancer and blood disorders, and their parents, to investigate their experiences with MyChart, a tethered PHR system. Through analyses of usage logs and independently-conducted surveys and interviews, we found that patients and parents both valued MyChart, but had different views about the role of the PHR for care communication and management, and different attitudes about its impact on the patient’s ability to manage care. Specific motivations for using MyChart included patient–parent coordination of care activities, communication around hospital encounters, and support for transitioning to adult care. Finally, some parents had concerns about certain diagnostic test results being made available to their children.

1. Introduction

In a pediatric cancer care setting, both adolescent patients and family members play an important role in the patient’s care management. Effective communication among adolescent patients, their parents and clinicians has proven to increase the quality and overall satisfaction of the care.¹² However, patients’ limited participation in the clinical setting, due in part to their limited health literacy, communication skills and perceived level of confidentiality, could lower the quality of communication.³

Personal Health Records (PHRs) show promising opportunities to alleviate these concerns, yet most currently do not provide different experiences for pediatric patients and their caregivers.⁴ Moreover, we do not fully understand the different information and communication needs of adolescents and their parental caregivers as they relate to the management of personal health records. Research examining how adolescents and their parents use PHRs has largely been limited due to federal and state regulations governing pediatric access to these systems.

In recent efforts to make electronic health records accessible to all patients, some states now offer patient portal enrollment to adolescents ages 12 and up, with proxy access available to their parental caregivers or legal guardians.⁵ Our research is motivated by the opportunity to investigate adolescents’ and parental caregivers’ experiences and actual use of the patient portal MyChart, a tethered PHR released by the Children’s Healthcare of Atlanta (CHOA). Understanding adolescents’ and parents’ attitudes toward and experiences with the portal—and investigating where they align and where they diverge—can shed light on the electronic information and communication needs of these consumer groups.

We present results of the first study documenting adolescents’ and caregivers’ actual use and reported experiences with a PHR system, along with the first empirical analysis of adolescent patient versus parental caregiver perspectives stemming from their experience with the system. As such, it is guided by the following research questions: Do adolescent pediatric patients and their parental caregivers use a tethered PHR system when it is available? What features of the PHR do they find valuable? How do patients resolve questions about its content? How does the tethered model of control impact these patients’ and parents’ perceptions of usefulness and expectations of privacy? Findings from our study contribute knowledge that can help guide the design of health information technology aimed at supporting adolescent patients with cancer and blood disorders and their parental caregivers.

2. Background

Research aimed at supporting patient access to clinical records through electronic patient portals has gained recently momentum.⁶ For adolescents undergoing frequent hospitalizations and complicated therapies, health IT features prominently in illness management and care. Many state regulations mandate that once a pediatric patient turns 13, parent access to their online record be deactivated. For children ages 13-17, the minor patient must authorize proxy

628
access for the adult. Once the patient turns 18, the now-adult patient must again grant proxy access through an authorization process. In order to make an informed choice to grant proxy access, the patient must understand what such access entails.

Protection of adolescents’ confidentiality in relation to their proxy has been a recent topic of interest in health informatics, with important medical, social and legal implications. Medical communities are aware of the implications of making electronic medical records accessible to minor patients: special requirements and challenges have been outlined, with emphasis on issues concerning the patient’s privacy and their access to sensitive health information. Concerns over lack of perceived confidentiality may deter adolescents from seeking medical care, including consultations with their doctors. Indeed, patients’ development of self-care skills and achievement of autonomy is critical for long-term outcomes. The Society for Adolescent Health and Medicine expressed this viewpoint, stating that “confidentiality protection is an essential component of health care for adolescents because it is consistent with the development of maturity and autonomy and without it adolescents will forego care.”

In theory, a PHR could alleviate some of these concerns by providing different viewing experiences for adolescent patients and adult caregivers. Yet, efforts toward designing for adolescents are still in their early phases. Ongoing efforts to understand adolescents’ attitudes toward health IT reveal tensions in information and communication needs of adolescent patients and their parents. For example, a focus group study with adolescents in a pediatric primary care setting found that adolescents had concerns about a lack of confidentiality of their communication, whereas parents were more concerned about “being left out of the loop,” or not being informed about significant health issues. Realizing the importance of privacy in adolescents’ care, recent efforts are focused on creating personally controlled health records (PCHR) that promise accessibility at the level of controllable individual features tailored to each patient’s needs. Yet, understanding which communications related to adolescent health require confidentiality is a particularly complicated problem, making both individual and hybrid models of control challenging to implement, particularly for complex illnesses that require family involvement.

Adolescents and young adults are known as being highly receptive to Internet search and mobile technology and are often early adopters of computing applications. For example, a recent study of adolescents’ actual use of smartphones revealed that, among many other types of applications, they used an average of ten distinct communication applications during two-thirds of the observed three hour period. A survey study exploring adolescents’ health information needs has shown that they typically searched the Internet to meet these needs. While adolescents are shown to use the Internet to a great extent, they still consider parents as their primary source for health-related information, reporting a twofold increase in satisfaction with parent-delivered information over Internet-acquired information.

Audit log studies on the actual use of patient portals can provide an empirical complement to self-reported attitudes and reveal gaps in uptake. Such analyses report disproportionate enrollment and use of patient portals by certain demographic populations, particularly including white, adult patients who are healthy and without Medicaid. While patient portals are now available to many, issues of information complexity and usability can hinder their adoption. To our knowledge, these difficulties have only been revealed in studies of adult use of PHRs. One study found socio-demographic disparities in their analyses of portal registrations among pediatric patients: the portal enrollment rate was lower for adolescents (12 years and older) as compared to infants and children (0 to 12) for whom their parents were predominantly involved in the enrollment and activation. Still, no studies of which we are aware investigate adolescents’ and caregivers’ ongoing experiences accessing a PHR system.

3. Methods

This work is part of a larger project on health information management practices of adolescents with cancer and blood disorders and their parental caregivers. The full study is an IRB-approved, multi-year and multi-phased project, which began in September 2014 in collaboration with IT staff and clinicians at Children’s Healthcare of Atlanta (CHOA). This paper reports results from a longitudinal exploration of patient and parent experiences with and attitudes toward the CHOA MyChart portal.

3.1. Study Site

Patients and parents were recruited in pairs at CHOA, a tertiary pediatric hospital in Georgia. We recruited participants in two Aflac Cancer and Blood Disorders Centers: Scottish Rite (suburban) and Egleston (urban), with each site serving different demographic populations. State regulations mandate that once a pediatric patient turns 13, parent access to their online record be deactivated. For children ages 13-17, the minor patient’s parent must consent to
allowing access, while their child must authorize proxy access for the adult. Once the patient turns 18, the now-adult patient must again grant proxy access through an authorization process. Our recruitment targeted a convenience sample of patient-parent pairs in the Aflac clinics, and was conducted both in-person and through flyers and email. We provided informed consent and assent forms to patient-parent pairs who met the eligibility criteria and guided each pair through the documents.

3.2. CHOA MyChart

CHOA released MyChart,1 a secure, HIPAA-compliant, tethered PHR in mid-summer 2014. Along with secure messaging capabilities to facilitate asynchronous electronic messaging between parents, pediatric patients and CHOA physicians, the tethered PHR (also referred to as a “patient portal”) includes access to laboratory test results, medication lists, patient allergies, prescription refill functions, appointment scheduling, messaging with clinical staff and the ability to store personal data. Once registered, patients and parents are given separate accounts with which they can access MyChart. The portal does not by default provide different viewing experiences of the patient’s information depending on the user, though proxies can be linked to several patients (their children). Physicians can customize whether messages should go directly to them (rather than being triaged first), to whom their message should go (to the proxy only or to both proxy and child), and whether or not an electronic reply to their message is allowed.

3.3. Study Design

To uncover potential attitudinal and experiential differences between adolescent patients and parental caregivers, we conducted a mixed-methods study comprising: 1) portal usage analysis, 2) small-scale survey, and 3) short interviews (see Figure 1). The usage analysis and interviews were conducted to contextualize survey responses.

Portal Usage Analysis

CHOA IT staff verified patient identification and registration and provided audit log data for patient and proxy usage over the study period. Auditing methods enabled us to examine information access at the level of individual access events (e.g., login events, loading of individual features, composing and sending electronic messages). Only patient log data (i.e., not its contents) were accessible to the research team. MyChart audit data included usage logs collected over a 19-month period ranging from August 2014 to February 2016. We analyzed the data focusing on commonly-used features, when they were accessed, and frequency of access over time.

Table 1. Excerpt of patient survey with selected questions.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Category</th>
<th>Example Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context of use</td>
<td>Closed-format</td>
<td>• “On which device did you access MyChart most?”</td>
</tr>
<tr>
<td>Preferred features</td>
<td>Open-format</td>
<td>• “What feature was most useful to you? Why?”</td>
</tr>
<tr>
<td>Experience viewing content</td>
<td>Likert-style</td>
<td>• “Viewing the information in MyChart makes me anxious.”</td>
</tr>
<tr>
<td>Communicating about health</td>
<td>Agreement</td>
<td>• “Some of the information in MyChart surprised me.”</td>
</tr>
<tr>
<td>topics</td>
<td></td>
<td>• “I asked someone in my family a question about information in MyChart.”</td>
</tr>
<tr>
<td>Role of PHR</td>
<td>Closed-format</td>
<td>• “MyChart would increase my engagement in my healthcare”</td>
</tr>
<tr>
<td></td>
<td>Yes/No</td>
<td>• “MyChart led me to ask questions that I might not have known to ask before”</td>
</tr>
<tr>
<td>Reasons to use PHR</td>
<td>Likert-style</td>
<td>• “When talking about something with my parents.”</td>
</tr>
<tr>
<td></td>
<td>Agreement</td>
<td>• “When speaking with another doctor about my care.”</td>
</tr>
</tbody>
</table>

Survey

Our survey instrument included a mix of open- and closed-format questions and several five-point Likert-style questions with scales designed to capture attitudes and preferences. Table 1 shows major topics covered, with selected questions. The survey took about 30 minutes to complete. Questions were matched for patients and parents, with minor changes to wording to make sure each question adequately addressed the participant type. For example, parents were shown statements such as “after using MyChart, I feel like I know more about my child's health,” whereas patients saw, “after using MyChart, I feel that I know more about my health.” We used REDCap,23 a HIPAA-compliant online research tool, to deploy the survey to patients and parents individually, with individual invitations sent after confirming that participants had used MyChart for at least one month.

---

1 http://www.choa.org/mychart
Interviews
The study concluded with a phone interview, conducted with adolescent patients and parental caregivers individually, once they completed the survey. Each interview lasted about 15 minutes and focused on confirming participants’ responses and eliciting elaboration on open-format responses.

4. Results
A total of 46 participants enrolled in our study including assenting patients (n=23) within the ages of 13 to 17 and their consenting parents (n=23). All recruited patients had been diagnosed with cancer or blood-related disorders. After collecting consent from all 23 parents and assent from all 23 adolescents, we introduced them to MyChart, explained data auditing plans and survey participation. Patients and parents were given separate accounts for logging in. We recruited additional study participants by sending recruitment ads to patients and parents who were already registered with MyChart. Each individual participant received a $25 gift card ($50 per pair) as a gratuity upon completing the study.

Some patient-parent pairs failed to complete MyChart activation. Of the 23 pairs, 12 (52.5%) patients and 15 caregivers (65.2%) responded to the survey. While 12 patients responded to the survey, only ten patients (mean age=15.3; male=3; female=7) and 15 parents (mean age=43.3) completed it. Survey responses were excluded if they were incomplete, or if we learned that someone other than the intended participant filled it out. In the following paragraphs, we report on survey results and audit log analyses for only those participants who completed the survey (Table 2 details patient and parent demographics).

We analyzed survey results and portal usage data using descriptive statistics. To explore similarities and differences in Likert-style survey responses between patient and parent groups, we report the difference in mean (mdiff) between the two groups, along with standard deviation (SD) values, for questions yielding the highest and lowest mean difference between the two groups. Below, we include verbal explanations of which group had higher or lower scores to accompany the mdiff value, which is reported as an absolute value.

We analyzed participants’ interview data through inductive coding to identify relevant themes in an iterative fashion. We organize our findings under three themes: perceived value of PHRs, keeping track of patient’s health, and electronic communication and sharing preferences. For each, we discuss patient and parent viewpoints, drawing attention to mean scores yielding the smallest and largest differences. Below, we refer to adolescent patients and parental caregivers as “patients” and “parents”, or T# and P#, respectively.

![Flowchart of participant enrollment in each study phase.](image-url)
4.1. Shared Perceived Value of PHRs

Analysis of survey responses with the smallest difference in mean showed that patients and parents both perceived MyChart as valuable. Both saw value in using the portal immediately before and after a visit to the doctor’s office. Both saw the most value in using it to navigate the transition from pediatric to adult care. Finally, they agreed strongly about situations when they would not use PHRs. Below we elaborate on each of these scenarios.

Before the visit: reminding and preparing for patient-provider encounters

Before a clinical encounter, patients (mean = 4.4; SD = 0.84) and parents (mean = 4.4; SD = 0.91) indicated that reminders were a valuable feature that allowed them to confirm their upcoming schedule, including future tests and doctor appointments, and to coordinate visits among family members. P12 told us, “I check it when he has an appointment. Sometimes I forget the exact time so I can always go in there. I usually check in a day or two before he has an appointment.” In the survey, T15 said that having the calendar reminder feature helps her make plans for other activities: “so I can plan my day and week and not forget times”

Patients and parents also saw utility in using the portal to prepare for clinical consultations. Both patients (mean = 4.4; SD = 0.7) and parents (mean = 4.53; SD = 0.64) indicated that knowing the information in the medical record in advance, such as lab results or imaging scans, could help them when talking to their doctor. In the exit interview, P12 told us that having access to her child’s record allowed her to seek better understanding of the lab results and their implications by asking targeted questions to the doctor: “because sometimes I see numbers on there that I’m not familiar [with]. They haven’t really explained about [the numbers] so I would go in and ask them.”

Having access to previous records in advance was an important feature for those who lived far from the clinic. P13 lamented about the need to travel long distances in order to receive her child’s medical record. She remarked, “living several hours away from T13’s doctors, it’s sometimes difficult to get info right away or even get in touch with someone who can give us any information. With MyChart, lab results go in, we can view them and decide what we need to discuss with the doctor instead of viewing them while at the appointment and not having much time to look over everything and ask questions.”

After the visit: fact checking and updates to the record

After a visit from the doctor’s office, patients and parents indicated that they would use MyChart to check that electronic information was correct and review updates to the record. For example, P12 said, “I’m trying to see if there are any updates (her labs, or any notes that were added).” The ability to see results after each visit helped one teenage patient talk to his doctor. When asked if using MyChart changed the way he talked to the doctor, T12 responded: “yeah, a little. Like one day I was talking [about] how much it [test results/blood level] would drop for my final week of the chemo and he said it would drop a lot and it [did]–it dropped a lot.”

---

### Table 2. Patient and parent survey participant demographics. A total of 10 patients and 15 parents completed the survey. Patient and parent pairs have matching ID numbers. T=Patient, P=Parent, E=Egleston, SR=Scottish Rite

<table>
<thead>
<tr>
<th>P ID</th>
<th>P Sex</th>
<th>P Age</th>
<th>P Srvy</th>
<th>T ID</th>
<th>T Sex</th>
<th>T Age</th>
<th>T Srvy</th>
<th>Primary Diagnosis (stage)</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td>M</td>
<td>43</td>
<td>Y</td>
<td>T3</td>
<td>M</td>
<td>14</td>
<td>Y</td>
<td>Osteosarcoma (metastatic)</td>
<td>E</td>
</tr>
<tr>
<td>P4</td>
<td>F</td>
<td>48</td>
<td>Y</td>
<td>T4</td>
<td>F</td>
<td>15</td>
<td>N</td>
<td>Osteosarcoma (neoadjuvant chemo)</td>
<td>E</td>
</tr>
<tr>
<td>P5</td>
<td>F</td>
<td>37</td>
<td>Y</td>
<td>T5</td>
<td>F</td>
<td>16</td>
<td>Y</td>
<td>Alveolar soft-part sarcoma (metastatic)</td>
<td>E</td>
</tr>
<tr>
<td>P6</td>
<td>F</td>
<td>37</td>
<td>Y</td>
<td>T6</td>
<td>F</td>
<td>17</td>
<td>N</td>
<td>Clear cell sarcoma (stage 4 metastatic)</td>
<td>E</td>
</tr>
<tr>
<td>P7</td>
<td>F</td>
<td>34</td>
<td>Y</td>
<td>T7</td>
<td>F</td>
<td>15</td>
<td>Y</td>
<td>Osteosarcoma (remission)</td>
<td>E</td>
</tr>
<tr>
<td>P8</td>
<td>M</td>
<td>52</td>
<td>Y</td>
<td>T8</td>
<td>M</td>
<td>17</td>
<td>N</td>
<td>Metastatic testicular germ cell (remission)</td>
<td>E</td>
</tr>
<tr>
<td>P11</td>
<td>F</td>
<td>38</td>
<td>Y</td>
<td>T11</td>
<td>F</td>
<td>16</td>
<td>Y</td>
<td>Juvenile granulosa cell tumor (remission)</td>
<td>E</td>
</tr>
<tr>
<td>P12</td>
<td>F</td>
<td>36</td>
<td>Y</td>
<td>T12</td>
<td>M</td>
<td>16</td>
<td>Y</td>
<td>Germ cell tumor</td>
<td>E</td>
</tr>
<tr>
<td>P13</td>
<td>F</td>
<td>49</td>
<td>Y</td>
<td>T13</td>
<td>F</td>
<td>15</td>
<td>Y</td>
<td>Idiopathic thrombocytopenic purpura</td>
<td>SR</td>
</tr>
<tr>
<td>P14</td>
<td>F</td>
<td>47</td>
<td>N</td>
<td>T14</td>
<td>F</td>
<td>13</td>
<td>Y</td>
<td>Sickle cell disease</td>
<td>SR</td>
</tr>
<tr>
<td>P15</td>
<td>F</td>
<td>50</td>
<td>Y</td>
<td>T15</td>
<td>F</td>
<td>14</td>
<td>Y</td>
<td>Osteosarcoma (stage 2)</td>
<td>E</td>
</tr>
<tr>
<td>P17</td>
<td>F</td>
<td>32</td>
<td>Y</td>
<td>T17</td>
<td>M</td>
<td>17</td>
<td>N</td>
<td>Malignant fibrous histiocytoma (remission)</td>
<td>E</td>
</tr>
<tr>
<td>P19</td>
<td>F</td>
<td>52</td>
<td>Y</td>
<td>T19</td>
<td>F</td>
<td>15</td>
<td>N</td>
<td>Von Willebrand disease</td>
<td>SR</td>
</tr>
<tr>
<td>P20</td>
<td>F</td>
<td>56</td>
<td>Y</td>
<td>T20</td>
<td>M</td>
<td>15</td>
<td>Y</td>
<td>Liver sarcoma (remission)</td>
<td>SR</td>
</tr>
<tr>
<td>P21</td>
<td>M</td>
<td>41</td>
<td>Y</td>
<td>T21</td>
<td>M</td>
<td>14</td>
<td>Y</td>
<td>Sickle cell disease</td>
<td>E</td>
</tr>
<tr>
<td>P23</td>
<td>F</td>
<td>41</td>
<td>Y</td>
<td>T23</td>
<td>F</td>
<td>17</td>
<td>Y</td>
<td>Wegener’s granulomatosis</td>
<td>E</td>
</tr>
</tbody>
</table>

---

632
Participants especially appreciated the ability to verify information in the record and promptly receive updates. P11 commented about the means to verify lab results, noting “it saves time. Making sure that everything is accurate. And on top of that I can actually see what the doctor’s put in the notes and everything. They tell you but if you want to get thorough, you can actually see what's going on.” Similarly, T23 responded, “We can usually find out results faster instead of waiting until my next appointment or waiting for the doctor to call.” Having prompt updates of lab results also helped ease T7’s frustration about delays between visits, “Yes. It was easier so I didn't have to wait like a month to ask my doctor these questions.”

Long-term prospects: care transitions

Both patients and parents agreed to a great extent that MyChart could serve to support the transition from pediatric to adult care over the long term, as well as better support communication with doctors in the short term. Interview data with patients revealed why this might be the case.

Both P7 and T7 commented on an upcoming transition to a different provider. In this case, they both saw MyChart as an important tool to archive information about the patient’s illness to retrieve later. For example, T7 remarked, “I think maybe a little bit more about my cancer history and just overall surgeries or just what happened. I think it's just good for me to know. So in the future, (when) I end up going to a different doctor or anything, I kind of know what was done to me.”

Both patients and parents responded that it would be important to refer to the information in MyChart during transitions to adult care (mean=4.6; mdiff=0), as well as to remain in touch with their current doctor once transitioning out of pediatric care (parent mean=4.33; patient mean=4.4; mdiff=0.07), and when speaking with another doctor about their care (parent mean=4.47; patient mean=4.4; mdiff=0.07).

P7 especially considered the need for access to her daughter’s histories, imagining a future scenario when her child would be entering college: “I can speak to her what the historic information is and kind of compare to whatever is told to me currently in regards to [T7]. I mean she's there ready to go to school and if there is a need to see a doctor who doesn’t have her history it would be readily available for them or even for her.”

Unhelpful use cases of MyChart

Not all aspects of MyChart appeared useful to adolescent patients and parents. For example, while both patients (mean=4.4; SD=1.07) and parents (mean=4.6; SD=0.6) appreciated the ability to see doctor's instructions or notes in the patient’s record, they were hesitant about adding new information or their own notes to the record (patient mean=3.5; parent mean=3.4; SD (both)=1.35).

As P7 explained: “my notes are my part of own notes, but I feel like MyChart…there should just be professional notes—nurses and doctors. There shouldn’t be any intertwining, as far as my opinions or my interpretation of that […] I take those notes for me personally in my journal, my book—not solely relying on MyChart.”

For some, MyChart was only useful during stages of diagnosis when patients and their parents were having several encounters with the hospital. Once entering recovery stages and remission, patients saw less value in using MyChart as they did before. T12 expressed this point, “if something comes up, I would use MyChart. But for now, since I'm in the recovering stage, I don't plan to. Like when I got off chemo, I used it one more time and stopped (using MyChart).”

4.2. Keeping Track of Patient’s Health

Analysis of survey responses with the greatest difference in mean showed that patients and parents had somewhat different views about the impact of the PHR on the patient’s ability to manage care. We learned that both used MyChart to make sense of the patient’s illness and treatment process, but they still sought information from external resources to resolve unclear information. Furthermore, adolescents relied on parents to provide explanations of information that was unclear to them. Below, we elaborate on these findings and provide analysis of portal usage activity and most commonly accessed features.

Perceived ease of use of MyChart for managing care

Overall, parents reported having experienced more difficulty than patients when using MyChart to keep track of their child’s health. When asked whether keeping track of the patient’s health was difficult, parents showed mixed sentiment (mean=3.53; SD=1.51; mdiff=1.13). On the other hand, patients were more likely to respond that they
experienced less difficulty in keeping track of their own health \((\text{mean}=2.4; \text{SD}=1.51; \text{mdiff}=1.13)\). Parents were also more likely to report their desire that MyChart be designed differently \((\text{mean}=2.8; \text{SD}=1.37; \text{mdiff}=0.9)\).

When compared to parents, adolescents reported a slightly more positive attitude toward the impact of MyChart on their ability to manage their care. After using MyChart, they reported having known more about their health \((\text{mean}=4.1; \text{mdiff}=0.7)\) in general and the care their doctor provides \((\text{mean}=3.9; \text{mdiff}=0.5)\). They also reported that the information in MyChart led them to ask questions that they might not have known to ask before \((\text{mean}=4.2; \text{mdiff}=0.6)\), and had slightly higher expectations that MyChart would lead them to take actions to improve their health \((\text{mean}=3.8; \text{mdiff}=0.4)\).

**Perceived ability to make sense of illness and treatment**

Having access to digital records allowed participants to engage with and understand their health differently. In the interview, P12 reflected on her experiences both prior to and after using MyChart to make sense of her child’s health: “they would just give us a paper with the numbers, and the following two weeks we had to go back and they would give us another paper so...I had nothing to compare [the current results] to anything before and I didn’t know where we were actually standing...” Since having access to her child’s medical record, P12 shared one of her exciting moments about how MyChart helped her and T12 make sense of the patient’s improved health: “When they told us that he was free of cancer, we saw the levels...we saw how [drastic]...that the numbers changed [...] I showed him where it started, during the chemo cycles and when it ended.”

For many, the portal was not the only source of information. Patients and parents also reported that they sought information on the Internet to help make sense of the medical record and seek additional information not readily available in MyChart. Nine out of 10 patients (90%) and 13 out of 15 parents (86.7%) had searched the Internet in the past to clarify information the medical record that was unclear to them. Reasons for searching typically included understanding medical terminology, interpreting lab results and radiology reports, and effects of medication on chemotherapy. While the Internet served general information needs for many patients and parents, it also seemed limited to some. For instance, P23 responded in our survey, “When reading a copy of the path report, it’s very difficult to just Google a word or term and find its meaning in relation to my daughter’s situation. Finding the words isn’t difficult. Translating what it all means seems almost pointless.” In this case, both P23 and T23 reported using private support groups or blogs to seek more information relevant to the patient’s condition. T23 said that she would look at, “blogs that have other people with the same disease as me.” We saw these preferences reflected in patients and parents’ attitudes about searching the Internet; While most had searched for information in the past, only four out of 10 patients (40%) and eight out of 15 parents (53.3%) reported that they would search the web in the future to understand information in their medical record.

**Mode of access**

When accessing MyChart, adolescent patients reported a greater preference for using smartphones or tablet devices over other modes of access including laptop and desktop PC \((\text{mean}=4.2; \text{SD}=1.03; \text{mdiff}=0.8)\). This finding is consistent with other studies that report adolescents’ inclination to use their smartphone for Internet use.\(^{18}\)

### Table 3. Summary of portal activity data analysis \((n=16)\).

<table>
<thead>
<tr>
<th>Activity</th>
<th>Min</th>
<th>1st Q</th>
<th>Mean</th>
<th>Median</th>
<th>3rd Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mos. since signup</td>
<td>9</td>
<td>13.3</td>
<td>14.4</td>
<td>14.5</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>Mos. active (from signup to last recorded activity)</td>
<td>1</td>
<td>4.8</td>
<td>8.9</td>
<td>9.5</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>Mos. MyChart data was accessed since signup</td>
<td>1</td>
<td>3.3</td>
<td>5.9</td>
<td>5</td>
<td>6.8</td>
<td>19</td>
</tr>
<tr>
<td>Activity over observed period (%)</td>
<td>5.9</td>
<td>22.5</td>
<td>41.2</td>
<td>42.0</td>
<td>46.1</td>
<td>100</td>
</tr>
<tr>
<td>Activity over active period (%)</td>
<td>23.1</td>
<td>51.8</td>
<td>73.7</td>
<td>73.2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Average access attempts per mo. since signup</td>
<td>1.2</td>
<td>8.4</td>
<td>22.4</td>
<td>13.4</td>
<td>20.9</td>
<td>131.8</td>
</tr>
<tr>
<td>Average access attempts per mo. over active period</td>
<td>9</td>
<td>16.2</td>
<td>31.6</td>
<td>20.7</td>
<td>33.6</td>
<td>131.8</td>
</tr>
</tbody>
</table>

### Table 4. Frequently accessed MyChart features \((n=16)\)

<table>
<thead>
<tr>
<th>MyChart Features</th>
<th># Accesses; (n=3732) (%)</th>
<th># of patients accessed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab results</td>
<td>985 (26.4)</td>
<td>13 (81.25)</td>
</tr>
<tr>
<td>Messaging</td>
<td>763 (20.4)</td>
<td>12 (75)</td>
</tr>
<tr>
<td>Lab Test Order</td>
<td>723 (19.4)</td>
<td>14 (87.5)</td>
</tr>
<tr>
<td>Appointment Review</td>
<td>425 (11.4)</td>
<td>12 (75)</td>
</tr>
<tr>
<td>Problem List</td>
<td>186 (5.0)</td>
<td>13 (81.25)</td>
</tr>
<tr>
<td>Immunizations</td>
<td>168 (4.5)</td>
<td>12 (75)</td>
</tr>
<tr>
<td>Allergies</td>
<td>162 (4.3)</td>
<td>12 (75)</td>
</tr>
<tr>
<td>Health Maintenance</td>
<td>161 (4.3)</td>
<td>12 (75)</td>
</tr>
<tr>
<td>Medication</td>
<td>159 (4.3)</td>
<td>12 (75)</td>
</tr>
</tbody>
</table>
Features accessed
Our analysis of audit logs shows that patients most frequently accessed MyChart features in the following order: Lab Results (25.6%), Messaging (20.4%), Lab Tests (20.3%) and Appointment Review (11.9%). Frequent use of these features supports findings of our survey regarding how patients and parents utilized MyChart for clinical encounters.

Most patients we audited regularly accessed MyChart over the 19-month observation period. Table 3 and Table 4 summarize analyzed usage. Participant activation occurred on a rolling basis over the 19-month period and averaged 8.9 months between the time of activation and the last-recorded activity (mid-Feb 2016). Most had periods of inactivity lasting one month or more. Participants logged into MyChart at least once a month (making it an “active month”) for an average of 5.9 months. Only three patient participants stopped using MyChart after one to two months of use. While patients were actively using MyChart, the period of active use and number of access attempts varied greatly across patients, as exhibited by wide gap between median and maximum access attempts. For example, T23, the most active user in our studied sample, accessed MyChart every month for the entire observed period and her average monthly access attempts equaled 131.8.

4.3. Communication and Sharing Preferences
Analysis of closed-format and Likert-style survey responses revealed insights about how patients and parents prefer to communicate with each other, with clinicians, and other people about their health regarding MyChart. The responses revealed slightly different attitudes and preferences between patients and parents when communicating with clinicians.

Adolescent patients communicated less frequently with clinicians through MyChart than their parents: only one out of 10 patients (10%) reported using MyChart to communicate with clinicians, whereas five out of 15 parents (33.3%) did. Patients’ reluctance to communicate directly with clinicians aligned with their preferences for communicating about their health in general: only two out of 10 patients (20%) reported they talked to their clinical caregivers about their health the most—the remaining eight talked to their parents the most and preferred to take questions about their health to parents over doctors. Seven out of 15 parents (46.7%) reported that they talked to a doctor or nurse about their child’s health more than they do the child. These findings suggest that parents act as intermediaries between clinicians and their children, even when electronic records are accessible to all.

Preferences for communicating about patient’s health
Adolescent patients and parents also had different viewpoints about their intention to communicate the patient’s health status to others, as well as different desires to learn about others like them. When asked if they would like to be able to share their health information with someone else, patients (mean=4; SD=0.94) agreed slightly more than parents (mean=3.6; SD=1.18). They also reported slightly more interest than parents in seeing information about other people who have similar health conditions (mean=3.9; SD=0.99; mdiff=0.37).

Concerns about viewing information in MyChart
The adolescent patients we surveyed indicated having no concerns about what their parents would see in MyChart. When asked if there was information in MyChart that patients would not like their parents to see, all patients reported that they were not concerned. When probed to describe any concerns, most patients reported having none. T23 remarked, “there’s nothing in MyChart that I wouldn’t share with my mom.” The same was true for all parents when they were asked if there was information in MyChart that they would not like their child to see. When further probed, some parents, however, some did have concerns. For instance, P13 noted her concern for the possibility that her child might misinterpret results of the diagnostic test, being concerned about “medical details she might misunderstand or misinterpret that might get her concerned, anxious or upset.”

Two parents noted that information containing a negative test result would be concerning if it were accessible to the patient only. For example, P15 wanted to see the negative result firsthand in an effort to allay the impact it may have on her daughter: “anything that would appear to be negative, I would like to see it first so that I’m prepared for any questions my daughter may have. Soften the blow, so to speak.” In an interview, P7 expressed similar sentiment that she would prefer having the information relayed to her first or when T7 is also present. She also preferred that the information be delivered in a different setting than MyChart. “I would prefer for her to know firsthand through MyChart, if there was a negative result. I would want her to know eventually. I wouldn’t hide that from her. But the way that I would deliver it to her... I would probably want to have that information in a different setting than her actually seeing that on MyChart.... My personal preference would be that the information is relayed to me first or to both of us together, but never to [T7] by herself.”
These concerns were upheld by other survey responses. When asked whether viewing the information in MyChart would make the patient anxious, patients (mean = 3.2; SD = 1.14) tended to agree with the statement slightly more than their parents (mean = 2.73; SD = 1.16).

5. Discussion and Conclusion

Parents and patients reported the importance of using the PHR shortly before and shortly after their clinical encounters. Survey results revealed that parents made more use of MyChart messaging features. Still, patients expressed more confidence in managing their health when using MyChart. They sought information in MyChart, and more broadly, the Internet, to clarify information in their records. Yet, consistent with previous findings, most of the patients in our study regarded their parents as their primary information source for health-related information and preferred to take questions about their health to their parents over their doctors. As part of their gradual transition to adulthood, teenage adolescents reconcile who is an authority on their health. Maintaining adolescent patients’ confidentiality while also meeting caregivers’ needs poses challenge for designers of health IT systems. In Hong et al.’s study of the ways in which adolescents with cancer and blood disorders participate in their care, the authors found that adolescent patients faced challenges representing their interests in communications with their physicians and relied on their parents heavily to manage their personal health.

Innovations in health IT are predicted to play a critical role in decision-making about treatment choices, care continuity, and improved measurement of outcomes of clinical trials. To make PHR systems valuable to adolescent patients and their parents, system design efforts must take into account the need to reconcile differing assessments of illness-related measures, and different communication preferences of adolescents and parents. We believe that many opportunities exist for health IT systems to provide age-appropriate mechanisms for reviewing clinical health data, and reporting on health status, health care experiences, and quality of life.

In particular, our findings point to the importance of further research focused on role of PHRs in supporting parental caregivers in communicating with their children, and in facilitating coordination and communication with clinical caregivers. Studies addressing teenage adolescents’ communication needs and preferences for health-related information will be critical to aiding what is currently a challenging transition from pediatric to adult patient care, particularly in light of the challenges inherent in providing adolescent and family access to PHRs. These challenges include the need to formulate unique consent and privacy laws that align with state and institutional policies, provide content at appropriate health literacy levels, and maintain accountability in the disclosure of health information while meeting expectations of privacy.

Limitations

By studying experiences with a tethered PHR, our study did not focus on the intricacies of enabling access control by patients or parents. To activate their accounts, patients and parents in our study agreed to the release of shared information in MyChart. The experiences, attitudes, and preferences we distilled should be viewed with the understanding that access control lay in the hands of the institution (i.e., limiting the sharing of individual data types or adjusting access control was not possible). We included a relatively small number of participants and focused only on the experiences of patients with cancer and blood disorders and their parental caregivers. We did not find conditions under which adolescents with cancer and blood disorders preferred to limit information, but prior work suggests that situations related to reproductive health, sexually transmitted diseases, substance abuse, domestic violence, and psychiatric concerns will demand fine-grained privacy protections in order to make PHRs useful. Assembling attitudes toward the use of PHRs and preferred avenues for communicating (and limiting or sharing) electronic data for these situations is an important avenue for future work.

Acknowledgements

We acknowledge the help of CHOA IT staff members Claudia Scott, Tod David, Heather Marney and Joshua Perkins and the assistance of Lauren Long and Tiffany Housworth. This study was funded by NSF #1464214 and a seed grant awarded by the Institute for People and Technology (IPaT) at Georgia Tech.

References


Cognitive Errors in Reconciling Complex Medication Lists

Jan Horsky, PhD, Harley Z. Ramelson, MD

Brigham & Women’s Hospital, Boston, MA; Harvard Medical School, Boston, MA; Partners HealthCare, Inc., Boston, MA

Abstract

Discrepancies between multiple electronic versions of patient medication records contribute to adverse drug events. Regular reconciliation increases their accuracy but is often inadequately supported by EHRs. We evaluated two systems with conceptually different interface designs for their effectiveness in resolving discrepancies. Eleven clinicians reconciled a complex list of 16 medications using both EHRs in the same standardized scenario. Errors such as omissions to add or discontinue a drug or to update a dose were analyzed. Clinicians made three times as many errors working with an EHR with lists arranged in a single column than when using a system with side-by-side lists. Excessive cognitive effort and reliance on memory was likely a strong contributing factor for lower accuracy of reconciliation. As errors increase with task difficulty, evaluations of reconciliation tools need to focus on complex prescribing scenarios to accurately assess effectiveness, error rate and whether they reduce risk to patient safety.

Introduction

Adverse drug events cause 770,000 injuries or deaths and cost $1.56 to $5.6 billion each year in the United States. Discrepancies between medications documented in electronic health records (EHR) systems and the actual state of their use by patients contribute significantly to the rate of preventable iatrogenic injuries. Over 25% of inpatient medication errors are due to inaccurate medication lists while errors in prescription medication histories occur in up to 67% of cases. The Office of the National Coordinator for Health IT (ONC) made medication reconciliation a core objective of its Meaningful Use (MU) initiative in recognition of its importance in reducing the incidence of adverse drug events (ADE) and the Joint Commission declared it to be one of National Patient Safety Goals. There is broad consensus that drug therapy errors can be reduced by making a complete, verified account of all medications a patient is actively taking and comparing this true state to lists maintained in one or more EHR systems. These records need to be updated regularly, especially at the time of a hospital admission, discharge or transfer and periodically maintained during ambulatory care. The maintenance process usually involves comparing medication history to the history in institutional records and to current prescription orders, identifying and reconciling discrepancies and documenting changes for future encounters and for other caregivers. Discrepancies between multiple lists and divergence from the true state of actively used medications can have serious consequences such as prolonged periods of over- or under-treatment that can only be detected by active surveillance.

Routine reconciliation is becoming more demanding with the steady increase of medication therapy, polypharmacy and the expansion of specialist care that allows more caregivers to write prescriptions for one patient. A nationally representative survey showed that 39% of adults over the age of 65 use five or more medications regularly. There is a corresponding rise in the difficulty of reconciling ever longer and more numerous lists from different care institutions and larger teams of providers. Reconciliation is conceptually simple and has been shown to be effective in reducing discrepancies but many clinics and hospitals found it difficult to implement a robust and reliable process with effective electronic tools. A recent study that evaluated the rate of inconsistencies and errors in medication lists reported that pharmacists reconciling records in ambulatory EHRs found on average 3.4 clinically important discrepancies per patient record.

There is strong research and anecdotal evidence from practitioners that the process is not supported well by currently available EHRs and that many find their reconciliation modules difficult to work with and poorly suited for the task. For example, a recent report showed that although vendors have been increasingly adding this functionality to their products in order to achieve meaningful use certification, clinicians in more than a third of the surveyed hospitals continued to use a partially paper-based process at admission, discharge or at ambulatory encounters as they were generally dissatisfied with the provided electronic tools.
Human factors and usability evaluation methods have recently started being applied to identify specific design constraints and workflow problems that affect the quality of reconciliation.\textsuperscript{13} Prioritizing recognition over recall, for example, is a common practice used by designers to decrease extraneous cognitive effort.\textsuperscript{14,15} make interaction less difficult and human performance less error-prone. During recognition, more cues and related information stored in memory are activated than in free recall and therefore the target item (e.g., medication name) can be more easily accessed and used in working memory.\textsuperscript{16,17} Guidelines and recommended usability principles commonly advise to use this concept in interface design.\textsuperscript{18} It is of particular significance when clinicians compare medications on two lists for differences and similarities.

As the result of care by many clinicians at different locations, multiple versions of medication lists may be stored on several information systems. What makes reconciliation more difficult is not only the number of medications that need to be reviewed but also the complexity of therapeutic regimens and the extent of detailed dosing instructions that need to be correctly interpreted and updated. While the list length increases the time and effort needed to account for the presence and absence of specific medications, more subtle differences between the same drugs or their generic equivalents present on both lists, such as changes in dose, frequency, route or use instructions, require sustained and focused attention to identify and enter them correctly.\textsuperscript{19,20} Substitutions of drugs in the same therapeutic class also need to be carefully reviewed by a trained clinician to avoid duplicate therapy.

This intrinsic difficulty is compounded by the need to make multiple comparisons to one or more lists and to the true state often confirmed by the patient or a caregiver. Centers for Medicare and Medicaid Services (CMS) estimate that a patient with one chronic disease sees up to four different doctors while those with five or more see an average of 14 different physicians,\textsuperscript{21} requiring more frequent and complex reconciliation of clinical information between all providers. Centers for Disease Control and Prevention (CDC) reported that 37% of Americans over the age of 60 take five or more medications and that the average number rises from five to seven at the age of 85.\textsuperscript{22} Patients with more comorbid conditions then have a higher probability that their medication lists contain errors and are at increased risk for adverse drug events.\textsuperscript{23}

There is a paucity of rigorously designed studies comparing different medication reconciliation practices.\textsuperscript{24} While some studies have shown the benefits of medication reconciliation technology almost all were done with systems developed in academic institutions, not by major vendors.\textsuperscript{24} The MARQUIS study reported, for example, that implementing a vendor EHR more than doubled the rate of medication discrepancies in one hospital.\textsuperscript{25,26} While there are well-designed, commercially available stand-alone systems for reconciliation available they are sparsely used due to the additional expense and difficulties with integration into EHRs that place extra burden on providers.

The majority (60%) of medication histories are shorter than five entries and the effect of list complexity may therefore be obscured in studies that do not stratify observations and results by the difficulty of reconciliation as errors tend to cluster within the subset of long and complex lists.\textsuperscript{9} Poor performance of suboptimally designed tools and their effect on patient safety may therefore be under-reported and under-studied.

The objective of our study was to investigate the effect of two different design concepts of a reconciliation module on human performance and error during a cognitively complex task. We asked clinicians to reconcile identical, standardized sets of medication lists with two different systems and analyzed the rate and character of observed errors. Our findings may inform the design of interventions and electronic tools that effectively alleviate cognitive burden on clinicians during complex reconciliations and reduce the number of errors leading to adverse events.

**Methods**

This study was designed as a standardized task scenario in which 11 clinicians performed reconciliation of two electronic lists containing 16 medications each using two different EHR systems in an alternating counterbalanced order to control for a possible learning effect. There were ten identical (including dose and frequency) medication records common to both lists, six were recorded only in EHR 1 and another 6 only in EHR 2 (Table 1). Reconciliation therefore required participants to make 22 drug comparisons: verify the ten identical entries and make a decision about correct updates to the local EHR system (e.g., add, don’t add, discontinue) on the remaining twelve. The test administrator simulated a patient and answered questions about the veracity of each medication.
The scenario contained four discrepancies that required participants to make a clinical decision rather than simply verifying that a particular medication was being actively taken by the patient. For example, Lasix, present on both lists (Table 1), required visual verification that the entries were identical in all aspects. When working with EHR 1, there was no action required for Ambien that the patient confirmed to be actively taking; however, on EHR 2, the drug would need to be added. Clonidine PO had to be discontinued when reconciling from EHR 1 since the patient was no longer taking it but no action was required when reconciling from EHR 2.

Table 1. Correct decisions and actions for reconciliation

<table>
<thead>
<tr>
<th>Medication</th>
<th>List</th>
<th>EHR 1</th>
<th>EHR 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abacavir 300 mg tab 2 oral daily</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Adefovir 10 mg tab 1 oral every other day</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Klonopin (clonazepam) 0.5 mg tab 1 oral bid, prn</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Lamivudine 150 mg 1 oral daily</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Lasix (furosemide) 20 mg 1 po every other day</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>MiraLax (polyethylene glycol) 142 mg/ml 17g oral daily, PRN</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Novolog R (insulin regular) 100 unit/ml sc sliding scale.</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Prilosec (omeprazole) 40 mg 1 oral bid</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Raltegravir 400 mg tab 1 oral bid</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Zofran (ondansetron) 4 mg/5ml 5ml oral q8h</td>
<td>Both</td>
<td>Verify</td>
<td>Verify</td>
</tr>
<tr>
<td>Ambien (zolpidem) 5mg po at bedtime PRN</td>
<td>1 only</td>
<td>Keep</td>
<td>Add</td>
</tr>
<tr>
<td>Aspirin 81 mg tab 1 oral daily</td>
<td>1 only</td>
<td>Keep</td>
<td>Add</td>
</tr>
<tr>
<td>Clonidine 0.1 mg PO BID</td>
<td>1 only</td>
<td>D/C</td>
<td>Don't add</td>
</tr>
<tr>
<td>Crestor (rosuvastatin) 40mg po daily</td>
<td>1 only</td>
<td>D/C</td>
<td>Don't add</td>
</tr>
<tr>
<td>Lisinopril 40 mg 1 po daily</td>
<td>1 only</td>
<td>Keep</td>
<td>Add</td>
</tr>
<tr>
<td>Zoloft (sertraline) 100 mg 1 oral daily</td>
<td>1 only</td>
<td>Keep</td>
<td>Add</td>
</tr>
<tr>
<td>Amlodipine 10 mg 1 PO daily</td>
<td>2 only</td>
<td>Add</td>
<td>Keep</td>
</tr>
<tr>
<td>Clonidine TTS 0.1 mg 1 patch qweek</td>
<td>2 only</td>
<td>Add</td>
<td>Keep</td>
</tr>
<tr>
<td>Humulin 70-30 unit/ml as directed, 34 units qam, 24 units qpm</td>
<td>2 only</td>
<td>Add</td>
<td>Keep</td>
</tr>
<tr>
<td>Lipitor (atorvastatin) 40 mg 1 oral daily</td>
<td>2 only</td>
<td>Add</td>
<td>Keep</td>
</tr>
<tr>
<td>Tramadol 50 mg 1 oral bid PRN pain</td>
<td>2 only</td>
<td>Don't add</td>
<td>D/C</td>
</tr>
<tr>
<td>Zestril (Lisinopril) 20 MG By Mouth every day</td>
<td>2 only</td>
<td>Don't add</td>
<td>D/C</td>
</tr>
</tbody>
</table>

Key

<table>
<thead>
<tr>
<th>Key</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both</td>
<td>Medications identical in all aspects were entered on lists in EHR 1 and EHR 2.</td>
</tr>
<tr>
<td>1 only or 2 only</td>
<td>Medication entered in one EHR only.</td>
</tr>
<tr>
<td>Verify</td>
<td>Compare entries on both lists; no action required if identical.</td>
</tr>
<tr>
<td>Add</td>
<td>Enter medication into the local EHR (present only on the remote list, confirmed as taken).</td>
</tr>
<tr>
<td>Keep</td>
<td>No action required (medication present only on the local list, confirmed as taken).</td>
</tr>
<tr>
<td>D/C</td>
<td>Discontinue in the local EHR (medication present only on the local list, not being taken).</td>
</tr>
<tr>
<td>Don't add</td>
<td>No action required (medication present only on the remote list, not being taken).</td>
</tr>
</tbody>
</table>

Discordant information presented in the scenario is listed in Table 2. The correct state specifies information confirmed by the patient (test administrator) in response to a direct question by the participant. The actions required to bring the medication list to a true state is shown from the perspective of EHR 2 being the local system. Conversely, the actions would be mirrored when EHR 1 was the local system.
Table 2. Decisions and actions for medication list reconciliation between two EHR systems

<table>
<thead>
<tr>
<th>Type</th>
<th>Discrepancy</th>
<th>Correct state</th>
<th>Reconciliation actions on EHR 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Brand vs. generic, dose change</td>
<td>Dose increased from 20 to 40 mg.</td>
<td>Zestril 20 mg (discontinue)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lisinopril 40 mg (add)</td>
</tr>
<tr>
<td>B</td>
<td>Same medication, change of route</td>
<td>Clonidine 0.1 mg PO bid (not taking), Clonidine 0.1 mg qweek 1 TD patch.</td>
<td>Clonidine 0.1 mg patch (keep)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Clonidine 0.1 mg PO (don’t add)</td>
</tr>
<tr>
<td>C</td>
<td>Duplicate therapy, same-class statins</td>
<td>Rosuvastatin 40 mg PO qday (not taking), Atorvastatin 40 mg PO qday</td>
<td>Lipitor 40 mg PO qday (keep), Crestor 40 mg PO qday (don’t add)</td>
</tr>
<tr>
<td>D</td>
<td>Patient not taking</td>
<td>Tramadol 50 mg PO bid PRN</td>
<td>Tramadol 50 mg PO (discontinue)</td>
</tr>
</tbody>
</table>

The reconciliation module of EHR 1 had the two lists visible on a single screen, side-by-side in two columns. The lists were also subdivided into therapeutic class sections that were horizontally lined up for direct visual comparison of individual entries within each section. In contrast, EHR 2 was designed with both lists in the same column – the remote list above the local list. Due to their lengths, however, clinicians could only see one at a time as the other was off the screen and required manual scrolling to display. A green highlight would appear over a pair of identical drugs (brand and generic versions) in the two lists on a mouse-over. However, the highlighted drugs could have different doses. A schematic of the basic layout concepts is in Figure 1.

![Figure 1. Schematic representation of the two layout concepts](image)

Participants were ten physicians and one physician assistant who have worked with EHR 2 daily for at least 6 months, except for two clinicians who had 4 months of similar experience. Most performed reconciliations routinely at least several times a week. None had previously reconciled medications using EHR 1 as the module had only recently been developed and implemented. They were shown its function in a short (1 minute) demonstration before the task started. The selected sample of 11 participants reflects a commonly recommended size for formative usability evaluation studies.27,28
Interactions of participants with the modules were recorded and analyzed with Morae 3.3 suite. The number and type of errors in the reconciled list as well as the time to complete tasks were determined. Statistical significance on the difference in the number of errors was evaluated using a non-parametric t-test.

**Results**

Ten clinicians completed reconciliation using EHR 1 with an average of 0.37 errors per participant (range 0 to 2; group total 3). The mean completion time was 5.5 minutes (range 1.96 to 8.95 minutes). Seven clinicians completed the same task (22 comparisons and decisions) using EHR 2 with an average of 1.29 errors per participant (range 0 to 3 errors; group total 9). The mean completion time was 7.8 minutes (range 3.7 to 9.6 minutes). Difference for the mean number of errors between systems approached significance at p≤ 0.057. Participants made between zero and three errors individually in each task (Figure 2).

![Figure 2. Distribution of the number of errors per participant](image)

Seven of the 12 errors observed during both variants of the task had similar characteristics and were aligned with discrepancy types A and B (from Table 2) and two with type C. Descriptions of the errors are in Table 3.

**Table 3. Type and frequency of observed errors**

<table>
<thead>
<tr>
<th>Type</th>
<th>Discrepancy</th>
<th>EHR 1</th>
<th>EHR 2</th>
<th>Comments (number of instances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Brand vs. generic, dose change</td>
<td>0</td>
<td>4</td>
<td>Both entries were discontinued (2). Entries were not verified for currently active dose (2).</td>
</tr>
<tr>
<td>B</td>
<td>Same medication, change of route</td>
<td>1</td>
<td>2</td>
<td>PO drug not discontinued – duplicate therapy (2). Wrong route (1)</td>
</tr>
<tr>
<td>C</td>
<td>Duplicate therapy, same-class statins</td>
<td>0</td>
<td>2</td>
<td>Crestor not discontinued – duplicate therapy (2).</td>
</tr>
<tr>
<td>Other</td>
<td>Ambien Humulin (insulin) Zoloft</td>
<td>2</td>
<td>1</td>
<td>Failed to add missing drugs to local list (3).</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
**Observed strategies and workarounds**

Clinicians working with EHR 2 used several strategies to compensate for the inability to see drugs on both lists at the same time for comparison. The most common approach was to scroll up and down for each drug: they asked the patient to verify a medication as being taken, located a matching entry on the local list (below) and updated or added as necessary. Alternatively, they would go through the remote list first in its entirety, added medications not identified by a green highlight as matched on the local list, reviewed the updated local list with the patient and adjusted the dose or discontinued the drug. Additional scrolling back to the remote list (above) was often necessary to compare doses, additional instructions and start dates. Clinicians recalled from memory which drugs have been already confirmed to speed up the process and to check whether there was a dose, brand-to-generic or route change.

Another strategy was to remove medications not being taken from both lists first and then review them one-by-one for differences, again scrolling for each comparison. Two clinicians explained that after a consolidation of the two lists (i.e., adding all drugs from the remote lists that did not highlight as a “match”) they left the module and completed the task using the EHR medication list screen.

**Discussion**

We evaluated two EHRs with differently conceptualized modules for medication reconciliation using a complex scenario and found a clear trend for higher accuracy and shorter completion time for one system. The key design difference was in the form of visual presentation of medication lists. EHR 1 showed lists side-by-side and organized all medications into subsections by therapeutic class that were horizontally lined up for direct visual comparison. EHR 2 presented the lists one above the other in a single column layout, with green highlighting of medications matched on both lists when the mouse hovered over an entry.

The effectiveness of appropriate cognitive support for the comparison task was evident in better performance of clinicians using EHR 1 and was also underscored by the fact that the participants had no prior training or experience with the module except for a short demonstration. In contrast, most participants worked daily with EHR 2 for six months or more and reconciled medications at least several times per week even at places where medical assistants and nurses performed the task routinely.

The most frequently observed error was a dosing error. The dose increase in the scenario was often not correctly interpreted and the locally recorded drug was either discontinued and the remote one not added or the dose difference was noticed and records were not compared to verify which one was active. Clinicians revealed in debriefings that those who relied on the green highlight function of EHR 2 did not realize that only the generic ingredient and route were checked automatically, not the dose. Clinicians who did not rely on the highlighting compared the entries directly and verified the right dose with the patient, as they did when using EHR 1. The ensuing medical error would probably result in therapy with an incorrect dose or no treatment of the condition for which it was prescribed. Although mandatory training for EHR 2 likely covered the meaning and function of the highlight, several participants were not aware of it and at least two assumed that a complete match included the dose was indicated. Problems and errors due to screen artifacts and visual cues that are not intuitive or may be misleading are rarely avoided by mandating more complete and thorough training, especially for very complex software where learning and skill develops over time with repeated use and under real work conditions.

Errors where there was a difference in route were observed on both systems. In EHR 2, clinicians looking at the second list had to recall correctly the entire medication instructions from the first list that was scrolled off the screen. Two participants made an error by recalling only the drug and dose and not the route and another one discovered the discrepancy only on a second review and corrected it. Several clinicians working with EHR 1 commented on the fact that the patch version of clonidine was placed in a different drug class section than the oral version, confounding the visual comparison of drugs within categories.

Duplicate therapy errors occurred only when participants worked with EHR 2. Clinicians correctly identified Lipitor as missing from the local list and added it but two failed to discontinue the existing Crestor. The duplicate therapy alert was not triggered at that point, allowing the possibility of an adverse event due to duplicate statin therapy.
All clinicians correctly revised the list after asking the patient about taking Tramadol and therefore there were no errors where the patient was not taking a medication and the medication needed to be removed from the active list. The other three errors resulted from clinicians not noticing the absence of a medication on the local list and failed to add it from the remote list.

Workarounds and interaction practices not originally envisioned by designers often indicate a poor fit of a tool to the task.\(^3\) In safety-critical work, unintended actions and re-purposed screen artifacts (e.g., free-text entries with content other than expected) also mean that some safeguards may be circumvented and new errors may be introduced.\(^31,32\)

In debriefings and interviews at the end of each session, clinicians recalled strategies and workarounds they use to make the reconciliation easier and faster under the constraints of a typical daily workload. For example, at one clinic, medical assistants were hired to reconcile the lists for all physicians as part of that institution’s transition to a new system. Their work was checked by accuracy for pharmacists who compared a printed version of each reconciled list for all patients at the clinic to an electronic version on the screen. This practice has been reported by other places that found their electronic tools to be lacking in effectiveness and usability.\(^12,33\) However, as two participants noted, off-loading routinely reconciliation to medical assistants to save time before visits is always risky as staff without adequate medical training would “fail to see that treating with Atorvastatin and Crestor would be duplicate therapy” and that their work would have to be double-checked for errors by a physician or a pharmacist. One participant explained that “I usually take a piece of paper, write the dose and date down and check it like that, because the most recent one is probably current; I wish I could see them both and not to have to scroll.”

The extra demand on attention was apparent by repeated checking of the same drug pair and multiple scrolling, especially for medications with similar names such as Lamivudine, Clonidine or Lisinopril when the clinician also had to recall which drug has already been viewed. Duplication of effort and extra time to complete was a common result of workarounds: in the words of one physician, “I would first ask you (patient) what you take and then go back up the list again.” Participants commented that this is “too cumbersome, it would be much better if I could see them side-by-side,” or that “this is a nightmare, I can’t see both.”

The differences in layout represented higher (one column) and lower (two-column) demand on cognitive effort and attention of clinicians. The one-column design was clearly inferior to the two-column concept. It necessitated repeated memorization and recall of items that may have had only subtle differences (e.g., different dose) or similar names (e.g., clonazepam, clonidine), required more attention and produced errors of omission, substitution and other failures of memory.\(^15,34\) When the top list (remote system) had five or six entries, the bottom list of similar length was already partially off the screen and disappeared from view at 12 or 13 entries. The difficulty of comparison would therefore be a problem for a substantial portion of a typical set of patient records in an EHR.

The two-column list allowed perceptual judgments and recognition of items rather than the more difficult recall. Typically, this approach lowers cognitive complexity and produces more accurate results.\(^35,36\) Sectioning of a long list of entries into more manageable chunks by drug class also simplifies a difficult lookup task by transforming it into a series of easier comparisons within each category where there usually are no more than four or five entries that are easier to verify visually. One unintended effect of assigning several categories to one drug was that the same medication with different routes could be displayed in different sections. Clonidine oral and transdermal patch, in our scenario, were not matched into the same sections on both lists, causing at least one error and several comments on this discrepancy from clinicians who eventually corrected their mistakes.

Although the length of the lists in this scenario was not typical, complex regimens are not rare in ambulatory medicine. Patients requiring extensive medication therapy also tend to be sicker and more susceptible to adverse drug events. Usability testing requirements for EHRs (such as those required for ONC certification) should more prominently focus on more complex and longer lists to evaluate more accurately the effectiveness of a tested system in preventing reconciliation errors.

**Limitations**

This was small-scale pilot assessment that will be complemented with a larger study. We have identified a trend in the difference of means of reconciliation errors but the results would have a much more general validity with a
larger cohort. We also experienced several technical difficulties with recording test sessions at locations in several hospitals and clinics that produced gaps in analyses and therefore only seven and nine clinicians completed tasks on each tested system. As planned, each participant would have completed the same task on both systems. However, there was no systematic exclusion and therefore the effect on validity should be minimal.

Conclusion

We tested two electronic medication reconciliation tools designed with different conceptual understandings of cognitively complex tasks. One system, currently used daily by thousands of clinicians, produced triple the rate of errors observed on a system more deliberately designed with sound usability and cognitive engineering principles. There is growing evidence that there are significant differences in the way clinicians interact with an electronic reconciliation tool and that the amount of time required and the accuracy of documentation is dependent on the quality of the design and its usability characteristics. Examining the intersection of health information technology and patient safety with practical conceptual models can advance the EHR-enabled healthcare system towards the goal of improving patient safety.

Acknowledgments

We want to express our gratitude to all clinicians who volunteered their time and contributed their skills and insights to the study. Many thanks are also due to Dr. Elisabeth Drucker for her essential support and help.

References


Clinic Workflow Simulations using Secondary EHR Data

Michelle R. Hribar, PhD1, David Biermann, PhD2, Sarah Read-Brown3, Leah Reznick, MD3,
Lorinna Lombardi, MD4, Mansi Parikh, MD5, Winston Chamberlain, MD, PhD3,
Thomas R. Yackel, MD, MPH, MS2, Michael F. Chiang, MD, MA1,3

1Medical Informatics & Clinical Epidemiology, 2School of Medicine, and 3Ophthalmology
Oregon Health & Science University, Portland, OR

Abstract

Clinicians today face increased patient loads, decreased reimbursements and potential negative productivity impacts of using electronic health records (EHR), but have little guidance on how to improve clinic efficiency. Discrete event simulation models are powerful tools for evaluating clinical workflow and improving efficiency, particularly when they are built from secondary EHR timing data. The purpose of this study is to demonstrate that these simulation models can be used for resource allocation decision making as well as for evaluating novel scheduling strategies in outpatient ophthalmology clinics. Key findings from this study are that: 1) secondary use of EHR timestamp data in simulation models represents clinic workflow, 2) simulations provide insight into the best allocation of resources in a clinic, 3) simulations provide critical information for schedule creation and decision making by clinic managers, and 4) simulation models built from EHR data are potentially generalizable.

Introduction

Physicians today are pressured to see more patients in less time for less reimbursement due to persistent concerns about the accessibility and cost of healthcare.1,2 Furthermore, clinicians are concerned that the adoption of electronic health records (EHRs) has negatively impacted their productivity.3–5 For example, at Oregon Health & Science University (OHSU), which completed a successful EHR implementation in 2006 that received national publicity, ophthalmologists currently see 3-5% fewer patients than before EHR implementation and require >40% additional time for each patient encounter.6

Facing these pressures, clinicians lack guidance on how to improve their efficiency while dealing with increased patient loads and time requirements of EHR use. For example, ophthalmologists typically see 15-30 patients or more in a half-day session, utilize multiple exam rooms simultaneously, work with ancillary staff (e.g., technicians, ophthalmic photographers), and examine patients in multiple stages (e.g., before and after dilation of eyes, before and after ophthalmic imaging studies). This creates enormous challenges in workflow and scheduling, and large variability in operational approaches.7 Approaches toward improving the efficiency of clinical workflow using EHRs would have significant real-world impact.

Clinic workflow bottlenecks result when patients arrive and clinic resources (e.g. staff, exam rooms, and providers) are not available to serve them. This mismatch of arrivals and availability can be increased by ad-hoc scheduling protocols that increase patient wait time.8 Testing different scheduling strategies and resource allocation in real clinics is impractical, however, since patient and provider time is too valuable for experimentation. Empirical models of clinical processes using discrete event simulation (DES) can evaluate different clinic configurations effectively before implementing them in clinical settings. DES requires large amounts of workflow timing data, which is available as timestamped EHR data.9 DES has been used for quality improvement in healthcare and scheduling, but not using EHR and detailed workflow data.10–13

In this paper, we present the process and results of simulating four outpatient ophthalmology clinics at OHSU, through discrete event simulation using secondary EHR data. Ophthalmology is an ideal domain for these studies because it is a high-volume field that combines both medical and surgical practices. Our results show that simulations can provide insight into the benefits and drawbacks of adding or removing clinic staff and exam rooms, as well as strategies for improving patient scheduling.

Background: Discrete Event Simulation

Discrete event simulation is a method for analyzing processes with high variability. The processes are broken down into a series of discrete steps whose time requirements are represented by probability distributions rather than constant values. When a simulation model is executed, these distributions are sampled to produce a time spent in each step. Simulations are repeated many times to determine the average behavior of the system. Simulation
models cannot be used to solve for absolute optimality; instead they are used to evaluate different scenarios to determine relative behavior.

For example, a clinic workflow can be modeled as a sequence of steps, such as patient arrival, an initial exam by a nurse or staff member followed by a physician exam as shown in Figure 1. Each of these steps will take a different varying amount of time, so each needs its own probability distribution. Each exam requires the use of finite clinic resources—a nurse or a doctor and an exam room. When the simulation model runs, a time value is determined for each step by sampling its distribution. As the model runs and more patients arrive, it is possible to generate queues of waiting patients for each exam. This happens when patients arrive faster than the exams take and/or there are not enough nurses, doctors or exam rooms for all waiting patients. To determine the expected behavior of the model over time, the simulation must be repeatedly executed.

Discrete event simulations can be used to evaluate changes to the model. For the example of the outpatient clinic, simulations can determine the effect of increasing the number of exam rooms or changing the appointment scheduling, which will both affect the patient arrivals. Each of these changes can greatly impact metrics such as average patient wait time or number of patients seen in a given time period. Simulation provides an easy and rapid way to evaluate these changes without interfering with clinic operations.

Discrete event simulations are used to improve various healthcare operations from emergency department configurations to operating room scheduling. Outpatient clinics have used discrete event simulations for resource allocation decision making as well as scheduling improvements, but mostly focus on individual clinics and use limited data as the basis of their model—multiple days of time-motion studies. In our study, we use multiple years worth of EHR timing data to create models that more precisely represent the variability of clinic workflows. We also parameterize the model so that it may be used for multiple clinics.

Methods

This study was approved by the Institutional Review Board at Oregon Health & Science University (OHSU).

Study Environment

OHSU is a large academic medical center in Portland, Oregon. The ophthalmology department includes over 50 faculty providers, who perform over 90,000 annual outpatient examinations. The department provides primary eye care, and serves as a major referral center in Pacific Northwest and nationally. We selected 4 outpatient ophthalmology clinics to study: 1) pediatric ophthalmology (LR), 2) comprehensive eye care (LL), 3) glaucoma (MP), and cornea (WC). These 4 clinics represent the diversity of outpatient care in ophthalmology at OHSU.

Over several years, an institution-wide EHR system (EpicCare; Epic Systems, Madison, WI) was implemented throughout OHSU. This vendor develops software for mid-size and large medical practices, is a market share leader among large hospitals, and has implemented its EHR systems at over 200 hospital systems in the United States. In 2006, all ophthalmologists at OHSU began using this EHR. All ambulatory practice management, clinical documentation, order entry, medication prescribing, and billing tasks are performed using components of the EHR.

Workflow Modeling and Reference Data Collection

Interviews with staff and observations of each of the four clinics were performed to determine the basic patient flow. All four clinics had the same basic clinic flow as shown in Figure 2. Patients check in and wait to be seen. An ancillary staff member performs an initial exam in an exam room. At the end of this exam, the patient’s eyes may be dilated. If this is the case, the patient returns to the waiting room while waiting for the dilation to take effect—about 25 to 30 minutes. After the dilation, the patient is returned to an exam room and waits for the physician exam.
If the patient’s eyes were not dilated, the patient remains in the exam room and waits for the physician. While the patient is waiting for the physician exam, the staff member must complete the documentation of the initial exam before the physician can start the exam. After the physician’s exam, the patients check out and leave. Because OHSU is an academic institution, trainees (residents and fellows) may also examine the patient. Trainees’ exams, however, do not occur at regular points in the workflow. Because of this added complexity, we focus on workflows without trainees in this initial study.

Once the workflow was understood, we performed time-motion studies for 3 - 6 half-day sessions at each of the clinics. One to two observers recorded timestamps of physicians and staff as they entered and exited exam rooms; these timestamps were then processed later to determine the duration of time spent in exam rooms with patients. This observational timing data was used to validate the use of EHR timestamp data to represent workflows and will be used to validate the simulation models in this study.  

Simulation Model Parameters and Their EHR Sources

To build the models for all four clinics, we first had to determine the necessary parameters and the EHR sources for them. In a prior study, we identified sources of workflow data within the EHR. We use the clinical data warehouse and ophthalmology datamart for OHSU’s EHR (EpicCare; Epic Systems, Madison, WI). While these timestamps are specific to OHSU’s implementation in ophthalmology, comparable timestamps are available for other vendors, installations and specialties.

The parameters are based on the clinic workflow, clinic resources and clinic scheduling:

1. **Start and End of Patient Encounter:**
   - **Check in and check out timestamps.** These timestamps are available in the ophthalmology datamart.

2. **Start and End of Staff and Physician Exams:**
   - **Audit log timestamps.** Timestamps from the audit log can be used to represent the beginning and ending of individual exams during the course of the office visit, which can calculate the duration of these exams. Data about providers and workstations was also required to determine the proper context for the timestamps.

3. **Dilation rate:**
   - **Structured ophthalmology documentation form and ophthalmology datamart.** Eye dilation information is entered in the structured ophthalmology documentation form of the EHR and is available from the ophthalmology datamart.

4. **Staff Documentation time:**
   - **Audit log timestamps.** The ancillary staff members must complete their documentation of the initial exam before the provider can start their exam. Using the audit log data combined with user and workstation data, we could measure the documentation time by measuring the time spent on the staff workstations after the initial exam.

5. **Number of ancillary staff and exam rooms:**
   - **Interviews and audit log data.** We used a combination of interviews and analysis of EHR users to determine typical values of these parameters. For many clinics, these parameters could vary from day to day, depending on other providers’ clinics in their specialty who shared these resources.

6. **The number and length of the scheduling blocks and number of patients per block:**
   - **Interviews and clinic encounter data from ophthalmology datamart.** We were able to obtain scheduling templates for each of the four clinics that guided how patients are scheduled; however, clinics rarely follow this template for every clinic day. Urgent patients are added on, schedules are overbooked, patients cancel or do not show. We reviewed the encounter data from the ophthalmology datamart to determine representative schedules for each of the clinics.
7. **Patient arrival patterns: check in and appointment times from ophthalmology datamart.** While all the clinics have only scheduled patients (no walk-ins), patients do not usually arrive at their scheduled time. This variation can have a great impact on the number of patients at the clinic at a given time. Since patients are scheduled, we measured the difference in time from their scheduled appointment time and their check in time. For all clinics, the large majority of patients are early.

**EHR Dataset**
Once we identified the parameters and their sources, we gathered datasets of office encounters for each of the four clinics for two years: 2013 and 2014. As with any large dataset, there are some encounters with incomplete data—we used only those that had valid check in and check out times. Further, we focused only on encounters that did not use trainees, since our initial models did not include them, as mentioned previously. Next, we wanted to find clinic days that represented a “typical” half-day session. This would include patient visits that happened during regularly scheduled half-day clinic sessions without large numbers of no-shows or overbooks. Therefore, we restricted the data to include clinic sessions whose number of patients fell within the 1st and 3rd quartile of all clinic sessions’ volume. We plan to address the atypical clinic days with future models. Finally, for each clinic, we then separated the dataset into a training set for building the model (80% of the data) and a test set for validating the model (20% of the data), in order to avoid overfitting the model.

**Simulation Models & Validation**
We used Arena simulation software to build a model of each clinic’s workflow using the training dataset. We used the same basic model for all four clinics, but customized it for each. We validated the 4 clinics’ simulation models by comparing metrics (total average exam time and average wait time) to the EHR test dataset.

**Model Experiments**
Once the clinic models were validated, we used them to evaluate different clinic scenarios. Specifically, we tested the effect of adding clinic resources such as staff members and exam rooms. Additionally we tested various methods of scheduling strategies in an attempt to better manage the negative impact of patient encounter (visit) variability on the clinic flow. Previous studies have shown that scheduling longer encounters with higher variability at the end of the day helps reduce wait time. To simulate this scenario, we identified the patient encounters that were the longest 1/5 of all encounters in the two year dataset. We then tested schedules with varying numbers of long encounters as well as different placements of these long encounters.

**Results**

**EHR, Observed and Simulated Encounters**
Table 1 gives the number of encounters in the EHR, observed and simulated datasets. The table shows how the number of encounters decreased in the EHR dataset as we restricted the set to encounters with complete data, no trainees, and during a typical day (number of patients is in 1st – 3rd quartile of ½ day clinic volume). The EHR training dataset was used to determine the probability distributions used in the simulation model. The test data, observed data without trainees and the simulated data sets are those that are used for validation.

**Simulation Model Parameters**
Building the simulation models required generating probability distributions for the patient arrivals, staff exams, staff documentation and provider exams and determining values for clinic parameters, shown in Table 2. The probability distributions were fit to the EHR training data. We used the average dilation rate for each clinic to determine the likelihood that a patient’s eyes are dilated after the initial exam, but we used median values for the...
other parameters (number of staff, number of exam rooms, number of scheduling blocks, block length and total number of patients per half-day clinic session) since the model requires discrete values instead of continuous averages.

Table 2: Model Parameters per ½ Day Clinic Session. Parameter values for each of the four clinics that represent a typical ½ day session. Values were determined from audit log and ophthalmology datamart encounter information. All values are medians, except for the dilation rate, which is an average over all ½ day sessions in the clinic dataset.

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Dilation Rate</th>
<th># Staff</th>
<th># Exam Rooms</th>
<th># Schedule Blocks</th>
<th>Length (Minutes)</th>
<th># Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornea (WC)</td>
<td>10%</td>
<td>3</td>
<td>4</td>
<td>13</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Comprehensive (LL)</td>
<td>64%</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Glaucoma (MP)</td>
<td>7%</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Pediatrics (LR)</td>
<td>43%</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>15</td>
<td>13</td>
</tr>
</tbody>
</table>

Validation of Simulation Models

We used the parameters and distributions to build a simulation model for each clinic in Arena. We ran each model 1000 times and compared the patient wait time and total exam time to the averages of the EHR test data set and observed dataset; see Table 1 for the sizes of these datasets. Note: the simulated data set had very large n since we ran the models for 1000 half-day clinics to determine the long-term behavior of the system.

Validating the model is crucial for ensuring that the models are representative. The validation results, given in Table 3, show that the simulated wait and total exam times (staff + provider) are close—within 5.5 – 6.8% on average—to those measured from the EHR test data set. Three out of the four clinics’ models mean wait time and exam time were under 5% different from the EHR test data set. Only two differences of the means were statistically significant—Glaucoma (MP) wait times and Pediatric (LR) exam times.

Because of the differences in size between the datasets, statistical tests do not conclusively represent the similarity or difference between the distributions of datasets. Instead, we present visualizations of the datasets’ distributions for comparison. Figures 3 and 4 show the densities of the three different data sets for wait time and total exam time. The means are close in all the graphs and there is significant overlap for all the datasets in all the clinics, which indicates that the simulated models are representative of real clinic workflows. Because the simulation models have a large n and the plotted data is average clinic wait time instead of individual encounter wait time (a limitation of the simulation software), the plots of the simulated data have a much narrower distribution than the other 2 datasets.
Model Experiments

Once the models were validated, we are able to use them to experiment with different clinic configurations:

1. **Experiment 1: Varying number of staff and exam rooms**

Simulation models were used to determine the effect of changing the availability of resources. For our four clinic models, we investigated the impact of varying the number of available exam rooms and ancillary staff members. Clinic managers commonly make decisions regarding these resources, but do not have good data for supporting them. **Figure 5** plots the simulated patient wait time against number of exam rooms and number of staff members. For each experiment, the non-varying resource was held constant at the value given in **Table 2**. Patient wait time decreases as the number of exam rooms and staff members increases from 1 to 3, but levels off after that point.

---

**Figure 3:** Exam time densities of the three datasets: simulated, EHR test data and observed data. There is significant overlap of the three different datasets which indicated the simulated model is representative of the real clinic workflow.

**Figure 4:** Wait time densities of the three datasets: simulated, EHR test data and observed data. There is significant overlap of the three different datasets which indicated the simulated model is representative of the real clinic workflow.
Table 4: Experiment 2: Simulated Patient Wait Times for Different Number of Long Encounters per ½-Day Clinic Session. When a clinic session has a larger number of long encounters, the simulated average patient wait time increases and the reverse is true for fewer number of long encounters. Ideally, clinics should try to keep the number of long encounters steady per clinic session to avoid the variability in wait times.

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Minimum</th>
<th>Mean</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornea (WC)</td>
<td>0</td>
<td>26.9</td>
<td>33.6</td>
<td>9</td>
</tr>
<tr>
<td>Comprehensive (LI)</td>
<td>0</td>
<td>34.6</td>
<td>40.0</td>
<td>6</td>
</tr>
<tr>
<td>Glaucoma (MP)</td>
<td>0</td>
<td>20.3</td>
<td>28.1</td>
<td>8</td>
</tr>
<tr>
<td>Pediatrics (LR)</td>
<td>0</td>
<td>24.9</td>
<td>27.0</td>
<td>7</td>
</tr>
</tbody>
</table>

Since there is still only one physician in the clinic, adding more resources does not improve wait times since the single physician becomes the bottleneck.

2. Experiment 2: Impact of varying numbers of “long” encounters per clinic session

Simulation models were used to examine the impact of encounters that are “long” and have high variability since long encounters monopolize resources, cause delays and create queues of waiting patients. To investigate the impact that long encounters have on clinic wait times, we first defined long encounters to be the longest 20% of all encounters for a clinic in the two year dataset. On average, we would expect that 20% of patients in a half-day clinic session would fall into this category, but in reality, the number of long encounters per half-day clinic session may vary when schedules do not limit them. We looked at the datasets and determined minimum, maximum and median numbers of long encounters per clinic session for each provider’s ½ day clinic session. We then used the simulated models to measure the range in patient wait times for these differing numbers of long encounters per session. Table 4 shows these results. For all four provider clinics, the wait times increase dramatically as the number of long encounters per session increases. For example, the Cornea (WC) clinic’s wait time increases from 26.9 minutes when there are 0 long encounters to 61.9 minutes when there are 9 long encounters, which is an increase of over 100%. The other clinics are similar. Keeping the number of long encounters steady per clinic session to avoid the variability in wait times.

3. Experiment 3: Determining the optimal arrangement of encounters in a clinic session using competing metrics.

After having established above that keeping the number of “long” encounters steady over all clinic sessions reduces the variability of wait time, simulation models were used to investigate the best placement of long encounters in a clinic session. Anecdotally, we observed that schedulers often put encounters expected to be longer at the start of the session, so that the clinic does not run over time; however, this may worsen patient wait time.

To investigate the impact of long encounter placement on these two competing metrics, we ran simulations to determine the average patient wait time and clinic length for different placements of long encounters. Results are shown in Figure 6. Long encounters placed at block 1 are at the start of the clinic session; larger block numbers...
represent appointment slots closer to the end of the clinic session. The graphs show that mean patient wait time was minimized by scheduling the long encounters at the end of the session, but this has the effect of increasing clinic length. The converse is true—clinic length is minimized by scheduling the long encounters at the beginning of the session, but this also maximizes the mean patient wait time. These graphs and corresponding data can help clinic managers determine the best compromise between the two competing metrics.

Discussion

This study has the following key findings: 1) secondary use of EHR timestamp data in simulation models accurately represents clinic workflow, 2) simulations can provide insight into the best allocation of resources in a clinic, 3) simulations can provide critical information for schedule creation and decision making by clinic managers, and 4) simulation models built from EHR data can be generalizable.

1. Secondary use of EHR timestamp data in simulation models accurately represents clinic workflows. Typically, secondary uses of EHR data have been for clinical research, quality assurance, and public health, rather than for operational purposes. While emergency departments have used EHR timing data for tracking patients and quality assurance, our study focuses on using EHR data for modeling outpatient workflows and studying clinic resource allocations and scheduling strategies. Our study shows that the data needed to study workflow can be mined from the EHR and that it represents general trends of clinic workflow in simulation models—simulated average patient wait time and average exam time are within 5% of the EHR data for three out of 4 of the clinics. Patient flow is a concern for all areas of healthcare, in both inpatient and outpatient settings. As patients move through the stages of their care, bottlenecks occur at points where demand for finite resources (providers, beds, etc.) exceeds availability. Using EHR large-scale timestamp data provides models that accurately represent the variability and can be used to study various clinic configurations.

2. Simulations can provide insight into the best allocation of resources in a clinic. Adding resources to clinics can help reduce patient wait times, but only to a certain point—3 staff members and 3 exam rooms for all four of the clinics. As long as there is a single physician for the clinic, additional resources will cause this physician to become a bottleneck. Therefore, the decreasing marginal benefit of additional staff members or exam rooms should be carefully weighed against the costs of adding these resources.

3. Simulations can provide critical information for schedule creation and decision making by clinic managers. Limiting the number of long encounters per clinic session and placing them wisely can help reduce wait time and clinic length. Keeping the number of long encounters stable for all clinic sessions can help reduce variability in wait times from session to session. Further, the placement of the long encounters affects wait time and clinic length with opposing effects—the closer the long encounters are to the end of the clinic session, average wait times are decreased.
and clinic length is increased. The simulated data provides necessary information for clinic managers to decide best how to balance these two competing metrics. We are currently testing schedules in clinic that place long encounters about an hour from the end of the clinic session since this minimizes wait time without unduly increasing the length of the clinic. Preliminary results show that the schedule implementation is effectively minimizing patient wait time.

4. Simulation models built from EHR data can be generalizable. By creating models for multiple outpatient clinics, we show that this approach is not limited to an individual clinic. As long as the simulation parameters are modified for each clinic, the models will still be representative. Further, even though the four clinics we modeled were quite different in terms of their clinic parameters (exam distributions, number of staff, number of exam rooms, number of patients, etc.), they all displayed the same general trends in all of the experiments. This gives us confidence that these trends are accurate and generalizable to other clinics.

Limitations

There are several limitations to our study. First, in order to identify audit log entries that correspond to exams, we had to discern when a staff member or physician was using the EHR during a patient interaction. At OHSU, we have uniquely named workstations in each exam room, which makes identifying patient interactions easier. If laptops were used, it would be much more difficult to determine when a provider was with a patient versus charting in an office. Second, the EHR timestamps do not always capture time spent with patients when the staff or doctor is not using the EHR. While we can determine what times the providers are not using the EHR, we cannot pinpoint what they are doing at those times. Second, the difference in size between our datasets (namely, the small observed set) that we used for the simulation validation limits our ability to use formal statistical methods for comparison. We are currently investigating methods for improving our observed clinic data collection so that we can easily increase the size of this dataset. Third, we limited this initial study to encounters that did not include trainees because of the differences in how trainees interact with the workflow. We are expanding our models to include this important activity to determine its impact on resources and scheduling. Finally, scheduling long encounters wisely improves patient wait times, but requires predictions of encounter length. We have performed preliminary studies using the relationship between clinical and demographic factors and visit length using timing data from the EHR; we are continuing to study this area, including adding physician’s input.

Conclusion and Future Directions

Simulation models based on secondary EHR timestamp data can be powerful tools for improving clinic workflows. Multiple years worth of clinic data provides an accurate representation of workflow variability in the models, which allows them to accurately evaluate the impact of clinic changes, whether it be adding staff or exam rooms or investigating novel scheduling approaches. The multiple models indicate that observed trends are real and generalizable. This implies that secondary use of EHR timestamp data for simulations broadens EHRs from a repository of clinical data to a holistic tool for managing clinical workflow.

Acknowledgements

Supported by grants 1T15 LM007088, 1K99 LM012238, and P30EY0105072 from the National Institutes of Health, (Bethesda, MD) and unrestricted departmental support from Research to Prevent Blindness (New York, NY).

References


Visualization of Order Set Creation and Usage Patterns in Early Implementation Phases of an Electronic Health Record
Nathan C. Hulse, PhD,1,2, Jaehoon Lee, PhD,1,2, Tim Borgeson, MBA3
1Intermountain Healthcare, Salt Lake City, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; 3Cerner Corporation, Kansas City, MO;

Abstract
Robust order set catalogs are considered to be a vital part of a computerized physician order entry (CPOE) implementation. Tools and processes for building, localizing, and maintaining these content sets in a centralized repository are important in facilitating the knowledge management lifecycle. Collectively, these order sets represent a significant investment of effort and expertise in capturing and distributing best clinical practice throughout an enterprise. In order to address an important gap of understanding how order sets are both created and used in practice in a current EHR installation, we have developed tools to analyze how order sets are used and customized in clinical practice. In this paper, we present the capabilities of these tools. We further characterize early development patterns in our enterprise order set catalog in early phases of a system-wide vendor EHR rollout. We present data that show how personalized order sets (favorites) are authored and then used in clinical practice. We anticipate that this type of utility will provide useful insight and feedback for those tasked with content governance and maintenance in CPOE systems.

Introduction
Computerized physician order entry (CPOE) systems pose an interesting challenge in clinical knowledge management. Order sets, or standardized groupings of orders targeted to specific clinical conditions or scenarios, have been shown as an important part of user acceptance of CPOE. When physicians use order set templates, they often find personal benefit from the convenience of large groups of clinically-relevant orders being readily available for selection and customization. Organizations can also find benefit, often using order sets as a vehicle to drive standardization processes and reduce unnecessary variation in clinical care.

While electronic health record vendors provide tools and infrastructure for supporting computerized ordering processes, the population of knowledge in the form of order sets is often left to the institution itself. Typical implementations often involve a significant effort or investment in building (or purchasing), refining, localizing, and maintaining order set catalogs.1,2 These content libraries are designed to support routine procedures and common conditions treated by care teams in clinical domains.3 Previous research4-9 has detailed many of the challenges associated with knowledge maintenance in order set catalogs, including:

- Lack of personnel and resources to build order set libraries at the enterprise level
- Immature and poorly-adopted standards for order sets and orderable catalogs impede the effective sharing of order sets from one institution to the next
- Building and maintaining order set catalogs requires a sizable investment in time and resources
- Implementations often suffer lower uptake of order sets within CPOE due to gaps in content coverage in order sets libraries
- Conversion of purchased libraries to local practice requires refinement
- Clinical knowledge advances rapidly and the need to update and maintain content within order sets is constant

The various phases of the knowledge management lifecycle are depicted in Figure 1. In preparation for an enterprise installation of a new EHR and CPOE system, we identified that visibility in order set authoring and usage data would be a key gap in our knowledge management processes.10 Content authors were quick to indicate their desire for systems that present utilization data as feedback to assist in refining their order sets. Previous research has shown that automated processes for summarizing utilization data can be turned into reporting feedback loops that facilitate content change.11-13
**Background**

Intermountain Healthcare is a not-for-profit integrated healthcare delivery system based in Salt Lake City, Utah. It provides healthcare for the residents of the Intermountain West, including the entire state of Utah and parts of southern Idaho. Intermountain maintains a health insurance plan, 22 inpatient hospitals (including a children’s hospital, an obstetrical facility and a dedicated orthopedic hospital), more than 185 outpatient clinics, and 18 community clinics serving uninsured and low-income patients. Collectively, it provides primary and specialty care for nearly half of the residents of the state of Utah. Informatics teams provide support to clinical users in the form of clinical decision support, clinical knowledge management, content maintenance and software development for unmet clinical needs.

In late 2013, Intermountain Healthcare and Cerner Corporation announced a strategic partnership, including an enterprise rollout of the Cerner EHR platform, and a joint development effort aimed at delivering novel informatics solutions for healthcare organizations. Implementation of the system is still underway, with several sites in the enterprise currently live on the new software. One of the strategic priorities of Intermountain’s Cerner installation focuses on implementing and measuring utilization patterns of best practice care process modules into the system, in the form of order sets, decision support rules and care pathways. In order to address the need of understanding and reporting how knowledge content input into the system is actually being used we set about to build an interactive dashboard that illustrates the usage of order sets, personalized order sets derived from them (hereafter referred to as ‘favorites’) and how collective differences in the composition of favorites, in concert with utilization data can drive suggestions for optimizing order sets.

![Knowledge Management Lifecycle](image)

*Figure 1- Knowledge Management Lifecycle. In this cycle, content is created, distributed, and used within a clinical system. Detailed usage data is captured, transformed into specific feedback, and provided back to content owners to facilitate further refinement.*

**Methods**

Working with technical staff and data analysts, we extracted data from relevant tables in our production Cerner Millennium database environment and moved the data to Intermountain’s data warehouse. For the purposes of development, approximately 12 months of usage data were extracted (February 2015 to February 2016). This resulted in data derived from nearly 300,000 order set usage instances. Filtering this data to exclude data derived from technical staff and/or testing data reduced the data set by approximately 20%.

Following several sessions with clinical sponsors focused on design and requirements definition, we used Tableau software to prepare an interactive dashboard that presents visual depictions of order set usage. Figure 2 shows the main view from this dashboard, in which all order sets in the catalog are visible, ranked by overall usage, with two
main color depictions illustrating overall ordering volumes from the standard template, as well as the collective favorites that were derived from it.

Figure 2 - Main view of order set utilization dashboard, showing content, usage data, filters, and order volume derived from the main template and its associated favorites. This specific view is focused on order sets in the Critical Care (CC) domain.

Figure 3 presents a graphical view of overall usage of an order set over time, also with baseline template and favorite ordering volumes depicted. Dynamic filters allow the user to focus strictly upon specific regions, facilities, providers, roles, or timeframes of interest to the user. Drill-down capability allows the user to select portions of the display and view only the data that pertains to the selection itself.

Figure 4 shows the ‘phylogeny’ of an order set, including timestamps, authors, and difference counts across the main versions of the order set itself. It further shows detail about the ordering physicians that have taken time to create favorites against the template, including timestamps, names, counts and drilldown views into the changes made in the favorite, and data regarding the number of times in which the favorite was used. At a high level, it illustrates the branching and evolution that can occur in the main order set template and the favorites derived from it.
Figure 5 depicts a line-item view into the details of an order set. When set to show all versions, the visual cue of color blocks next to each of the orders demonstrates the presence or absence of an order from a particular version of an order set. When viewed against the entire version history of the same content (or against favorites derived from the template) the color patterns show the evolution of the content itself, including orders that are added, removed, or updated over time. The color of the blocks themselves represent selection status in this view, indicating whether the order was preselected or not. Numbers across the top show the various versions of the content itself, as well as corresponding ‘denominators’, which indicate the number of times in which that particular version of an order set was ordered against. Numbers overlaying the orderable blocks themselves show the frequency with which particular orders were ordered against in the ordering sessions that used these order sets.

---

Figure 4 - Version and Favorites Tree for the CC General ICU Admission Order Set from the Intermountain Healthcare order set catalog. This view shows the version history, favorites branches, as well as the corresponding timestamps, authors (names and roles), number of ordering instances, and modification counts derived from each knowledge artifact.
This dashboard utility and these specific visualizations have allowed us to compile useful information about the scope of our current order set catalog and the extent to which it has been revised and used. It has also given us new insight as to the frequency with which our users are using personalized order sets to find, tailor, and adjust the enterprise order sets themselves. The insight is useful both in comparisons of templates to their derived ‘favorites’ collections, as well as across versions of the enterprise order set template itself.

**Results**

Our efforts in extracting these data sets and consolidating them have allowed us to summarize and describe the artifacts in the order set catalog itself. Table 1 contains data pertaining to the main order set templates in the library itself, their usage to date (with the understanding that our enterprise installation is still in early phases), and the number of authors responsible for the content. It shows that less than half of the available enterprise order sets available in the order set catalog have been used to date.

**Table 1. Summary data for enterprise order set template data in Intermountain’s order set catalog**

<table>
<thead>
<tr>
<th>TEMPLATE DATA</th>
<th>Care Plan</th>
<th>Phased</th>
<th>Pathway</th>
<th>Future Encounter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order Sets available</td>
<td>2,516</td>
<td>4,366</td>
<td>758</td>
<td>4,838</td>
</tr>
<tr>
<td>Used – to date –</td>
<td>970 (39%)</td>
<td>1,725 (40%)</td>
<td>386 (51%)</td>
<td>390 (8%)</td>
</tr>
<tr>
<td>Not Used – to date –</td>
<td>1,546 (61%)</td>
<td>2,641 (60%)</td>
<td>372 (49%)</td>
<td>4,448 (92%)</td>
</tr>
<tr>
<td>Total # of Usage Instances</td>
<td>131,498</td>
<td>137,869</td>
<td>3,134</td>
<td>4,785</td>
</tr>
<tr>
<td>Grand Total of Usage Instances</td>
<td>270,118</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique Content Authors</td>
<td>107</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 summarizes similar data for favorites saved to the order set catalog as well as corresponding usage data. This table contains two fewer order set types, as users do not have the ability to set favorites on future (day of treatment) and phased order sets in the Cerner system. It also shows how many order set templates have been ‘favorited’ and the average number of favorites per ‘favorited order set’. This table also shows the number of ‘cloned’ favorite order sets, in which the user has personalized the order set in name only (the content of the personalized order set does not differ from that of the template from which it was derived).

**Table 2.** Summary data for personalized or ‘favorites’ order set data in Intermountain’s Cerner order set catalog

<table>
<thead>
<tr>
<th>FAVORITES DATA</th>
<th>Care Plan</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorite Order Sets available</td>
<td>8,787</td>
<td>6694</td>
</tr>
<tr>
<td>Derived from unique order sets</td>
<td>693</td>
<td>275</td>
</tr>
<tr>
<td>Average # of favorites per ‘favorited template’</td>
<td>12.7</td>
<td>24.3</td>
</tr>
<tr>
<td>Used – to date –</td>
<td>3282 (37%)</td>
<td>3307 (49%)</td>
</tr>
<tr>
<td>Not Used – to date –</td>
<td>5505 (63%)</td>
<td>3387 (51%)</td>
</tr>
<tr>
<td>Total # of Usage Instances</td>
<td>22,670</td>
<td>15,773</td>
</tr>
<tr>
<td># that are exact copies of template (clones)</td>
<td>954 (11%)</td>
<td>509 (8%)</td>
</tr>
<tr>
<td>Unique Content Authors</td>
<td>1424</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 depicts the additions to the collection over time as authors added to and updated existing content in the set. It shows additions to the order set catalog in the year prior to the initial go-live in Intermountain’s Cerner installation (February 2015). Favorites are shown in orange, while main templates and revisions are shown in blue.

![Figure 6 - Content additions to the Intermountain order set catalog (consisting of both enterprise order set templates in blue and the entire personal favorites catalog in orange) in 2014 through early 2016. Enterprise order sets and revisions are shown in blue, personalized order sets or 'favorites' are shown in orange.](image)

Figure 7 shows the overall volumes of orders derived from order sets coming into the database from the initial enterprise installation to present (February 2015 to February 2016). This graph, combined with the data shown in
Tables 1 and 2 show that despite the large numbers of favorites available in the Intermountain order set catalog, most of the order volume derived from CPOE is coming from enterprise template order sets, not favorites. Growth in the volumes shown in the order curve are derived largely from new sites and regions coming online with the new Cerner EHR system.

**Figure 7 - Order Set usage data since initial Cerner implementation at Intermountain Healthcare.** Order sets entered via 'favorites' are shown in orange, enterprise-level content is shown in blue.

**Discussion**

We have constructed tools that allow order set content owners to quickly identify their content, view its utilization over time, and understand specific details of how, where, and who is using the content. Further, material views make it relatively straightforward for content owners of order sets to understand how their content is being used. As we have developed and presented these views back to the clinical teams responsible for the content, they have expressed both interest and enthusiasm in being able to quickly visualize how and where their content is being used. Several key themes have emerged as we have reviewed this data and presented it back to users.

**Departmental differences over Favorites**

Different departments and clinical programs inside Intermountain Healthcare have strikingly different views on the place of personalized order sets within a CPOE system. Some clinical groups (like cardiology) are patently opposed to users utilizing them and have configured the system to disallow specific user roles from creating them. Other departments are much more relaxed in this regard and go as far as encouraging them and training users how and when to go about it. As a result of these differences, departmental level views of order set utilization can be very different, with cardiology-specific content showing almost no utilization of favorites, to the other extreme in
obstetrics/gynecology where over 70% of postpartum discharge orders are entered via favorites. Despite the difference in philosophy or approach on personalized order sets, both departments have expressed value in being able to visualize the usage patterns of their content.

**Namespacing or ‘initialing’ of content**

As shown in Table 2, between 10 and 15% of all favorites in the Intermountain catalog are exact copies of the enterprise order set from which they were derived, or ‘clones’. In many cases, the only material difference between these personalized order sets and the templates from which they were built is the name. Many users applied a namespacing approach in which they preprended or appended their initials to the title of the order set. Others renamed them to facilitate searching using terms they would think of when using the search mechanism of the Cerner CPOE system. Due to the updating features of the Cerner EHR system, these users’ content will remain consistent with the enterprise template as it evolves.

**Content gaps**

The phylogeny-based view of the order set content also revealed interesting perceived gaps in the order set catalog itself. Simple inspection of the obstetrics/gynecology order set data showed large amounts of favorite use in the postpartum discharge order sets. Upon inspection of the favorite sets built against this template, key themes emerged from the order set titles, in particular the groundswell of desire from the users to separate the content into separate groupings of ‘C-section discharge orders’ and normal or vaginal delivery C-section orders. When content authors originally built this order set, they had grouped these two use cases (post C-section discharge orders and normal vaginal delivery orders) into one order set intended to support both. Yet the users’ creation and subsequent use of these two separate ‘favorites-based’ sets indicates a strong preference for separate content to support both scenarios. This type of feedback allows content owners to revisit and potentially re-prioritize content development efforts.

**Unused content**

In both the enterprise order set and favorites libraries there were significant numbers of order set templates that have not yet been used. This is not entirely unsurprising, as Intermountain is still relatively early in its implementation of the Cerner EHR system. The significant number of unused ‘favorites’ is somewhat more difficult to characterize, though some of that may be due to content created by physicians during their training with the software or time with physician coaches who are orienting them to the features present in software that is unfamiliar to them.

Over time, this usage data may bring to bear content that is present in the library, but semi-obscetate from a utilization perspective. Seldom-used order sets may be considered for deletion from the library in such circumstances, so as to optimize the users experience with the software, as well as to minimize the already significant maintenance costs that come with keeping them current with best practice and standards given in the medical literature.

Some clinical groups have been a bit surprised by the utilization data, in that they have in some cases created content that is rarely used. Some clinical workflows and even training routines that are part of implementation favor other methods for ordering, including quick orders, workflow-driven wizard screens and other modules. In some of these cases, content owners have shifted some of their focus away from order set development for all scenarios, targeting instead those where they feel that other options don’t address the clinical users’ needs adequately.

**Ownership/governance**

Tables 1 and 2 illustrate the scope of authors required to create the order set library that we are characterizing. It details that there are just over one hundred content authors who maintain the enterprise order set library and well over 1400 individuals that have personalized order sets. At Intermountain, the order set creation and update effort is spread across clinical programs at the corporate level, with regional representation from the various sites/regions and input from clinical, administrative, and central groups (e.g. compliance, legal, etc). It also has input from clinical service lines, local units, physician champions and specialists. Clearly there are many other individuals beyond the 100+ specific content authors who are expending significant effort in curating/refining content in the enterprise content development teams. The overall effort necessary to build these libraries is significant.

**Order volume**

Figure 7 shows that most of the order volume derived from order sets does indeed come from enterprise-level order sets, and not from personalized ones. More than 80% of the volume collectively comes from the standard content.
There are clear exceptions to this, as noted above, and these differences may reflect different philosophical approaches in clinical divisions, content gaps and/or other reasons specific to the users. Our hope is that presenting these data to content stewards can stimulate meaningful conversation amongst authors and users of these knowledge artifacts to optimize users’ experience with the order sets.

Limitations
We have not looked at complete ordering patterns within the Cerner toolset, only those derived from order sets. Quick orders and other workflows support ordering and encompass the broader picture of medical orders in the clinical setting. Our analysis has not yet extended itself to differences in order sentence construction. We have also not yet undertaken the visual analysis of phased or component order sets which are common to multiple parent or pathway order sets.

Next Steps
We intend to extend the capabilities of the visualization toolset to drill down to the details of the orderable/order sentence level, further illustrating the specifics of how the orderables are being used. We also plan to work with clinical analysts from each of the clinical specialties to begin to attach cohort specifications as an applicable filter for viewing order set utilization data (e.g. to show usage of the febrile infant order set specifically for patients who met the febrile infant definitions from the clinical data).

Acknowledgments
We acknowledge the Cerner technical engineers who have played an important role in orienting us to the nuance of the clinical data used in the analysis. We likewise express appreciation for the support from Intermountain Healthcare’s enterprise data warehouse team. We thank Sameer Badlani, former CHIO at Intermountain, for his inspiration to begin the effort.

Conclusion
We have constructed tools that allow order set content owners to quickly identify their content, view its utilization over time, and understand specific details of how, where, and who is using the content. Materializing this information visually has illustrated some key themes of content usage, needed areas for refinement, and greater visibility into the actual consumption of knowledge content produced by central content teams. We anticipate that this vision into the data will lead to optimization and enhancement of content over time.

References
Discordant patient pain level reporting between questionnaires and physician encounters of the same day

David A. Juckett, PhD1, Fred N. Davis, MD2, Mark Gostine, MD3, Eric P. Kasten, PhD1, Philip L. Reed, PhD1, Joseph Gardiner, PhD1, Rebecca Risko, BSN2
1Michigan State University, East Lansing, MI; 2ProCare Systems, Inc., Grand Rapids, MI; 3Michigan Pain Consultants, LLC, Grand Rapids, MI

Abstract

We examined the consistency of pain reporting by patients in a community pain management practice in Michigan. We compared pain levels (range 0-10) entered by patients in questionnaires versus those provided during their face-to-face physician encounter on the same day. Both of these values were available for approximately 10,000 encounters during the study period (2010–2014). Two subpopulations of patients were identifiable. One was consistent in reporting worst or least pain levels on the questionnaire and during the provider encounter. The other was discordant. Factor analysis had previously identified severity scales for patient biopsychosocial characteristics derived from the full questionnaire. The two subpopulations differed in their factor profiles even though they had similar demographics. In general, pain reported directly to physicians was more correlated to biopsychosocial indicators. Pain self-reporting using questionnaires has often been assumed to be ground truth, but those obtained during the physician encounter may be more reliable.

Introduction

Chronic pain in the US is a massive medical problem and treating pain is a major challenge, as emphasized by the Institute of Medicine’s report on relieving pain in America1,2. That report offers several recommendations to address chronic pain, which include the need to better understand pain in the population and the need to offer tools to community physicians for pain management. There is a growing consensus that evidence-based medicine should be the guide for developing these tools and treatments, but this is challenging because the biopsychosocial nature of pain must be addressed and the linkages between diagnoses, treatments, and outcomes must be discerned on a scale to match the problem.

To move toward addressing these issues, there is a need to capture and analyze both the patient and provider perspective during the routine encounters of clinical practice4. Patient perspectives are critical because there is no instrumentation to quantify pain levels and the patient’s psychosocial attributes contribute to their condition. Provider perspectives are critical because chronic pain is a set of complex syndromes that requires years of experience to understand and an ability to interact with the patient in pain. Currently, the medical information from routine clinical encounters is locked away in patient charts. Extracting such content from community pain medicine clinics would constitute an opportunity for retrospective comparative effectiveness evaluations to help build an understanding of existing treatments and patient phenotypes within the community.

Toward that end, this work utilizes a data repository from the Michigan Pain Consultants (MPC) group, which has electronically captured both the patients’ and physicians’ perspectives from thousands of patient-visits in a multi-clinic community pain medicine practice in Western Michigan. Michigan State University (MSU) has partnered with MPC and their practice management group, ProCare Systems, Inc. (ProCare), to share data and expertise.

The first step in the analysis of clinical data is to evaluate the reporting of pain by the patient. Relaying pain levels to the care provider is fundamental to proper patient evaluation and its accuracy is intertwined with the patient’s physical and emotional state together with their desired goals5. Pain levels can be captured as structured data within questionnaires or EMR fields, and/or as unstructured data within physician notes summarizing discussions during the patient-provider encounter. We compared two of these reporting mechanisms and have discovered that some patients significantly alter their pain reporting in the presence of the physician. This behavior allowed the identification of patient subpopulations that differed in their questionnaire pain reporting and provider encounter reporting for the same pain query, on the same visit.

In this paper, we extract pain values from progress notes using natural language processing (NLP) approaches and examine their frequency distributions after sorting by pain values recorded in the patient pre-encounter
questionnaire. Patient subpopulations exhibiting concordant and discordant pain reporting were detected within distributions for worst pain, least pain, and today’s pain queries. These populations could be isolated for least and worst pain allowing comparisons to previously determined biopsychosocial factors derived from subsets of questionnaire items. We show that pain values reported during the physician encounter were more highly correlated to severity levels of the biopsychosocial factors than those pain values provided on the questionnaire in the waiting room just moments before the encounter. This phenomenon grew more robust during follow-up visits. It is an effect that would go undetected by examining questionnaires alone and indicates that interpretation of biomedical data in electronic health records must account for the manner in which it is collected and the potential psychological drivers influencing patient veracity. We briefly discuss these findings in light of the importance of both physician and patient reports of pain and the implications for data capture in pain research.

Data and Methods

Data

The partnership between MSU, MPC, and ProCare, and the nature of the questionnaire data has been described previously. The content of MPC’s clinical records comprises practice management data, results from detailed patient questionnaires, and detailed progress notes, all derived from approximately 95,000 patient visits per year. MSU access to the data was facilitated by a Business Associates Agreement and the study has received exempt status by the MSU IRB.

As part of MPC’s regular assessment, patients are asked to complete a pain health assessment questionnaire (PHA ™), which is a patient self-assessment instrument that provides demographic, medical and social history, as well as patient-reported outcomes (PRO). This has been described in detail previously. During the study period, a total of 22,691 intake questionnaires (iPHA) and 61,175 follow up or cumulative PHAs (cPHA) were completed by patients during their initial visit and selected repeat visits, respectively. Among the iPHA, 12,983 could be linked to physician notes for the same day and, similarly, 37,348 of the cPHA questionnaires could be linked to progress notes generated during the same follow-up visit. (Note: Not all visits entailed seeing the physician and not all physician encounters resulted in progress notes being generated.)

Analysis Overview

The overall outline of the steps in the data extraction and analyses is as follows: 1. Identifying pain in progress notes using NLP; 2. Collect progress note values into histograms sorted by PHA questionnaire pain values; 3. Decompose histograms into peaks representing two patient groups; 4. Identify concordant (sync) and discordant (nonsync) pain reporting patient subpopulations; and 5. Compare characteristics of the two subpopulations.

Progress Note Natural Language Processing (NLP)

Background. The dictated progress notes are in the form of letters to referring physicians. In prior work, a random sample of 500 clinical notes was drawn from the year 2010 (stratified by age, sex, number of ICD-9CM codes per patient, number of CPT codes per patient, and clinical note length), as described previously. Analyses of these notes revealed a highly organized structure generally following a SOAP note (Subjective, Objective, Assessment, Plan) structure although more comprehensive. The pattern typically comprises; reason for visit, pain status, past outcomes, history, findings, diagnosis, plan, procedure (if any), and current outcomes. The progress notes provide a rich source of information on patient status seldom captured in structured fields of EMRs. Our goal is to use this information to help construct phenotypes for patients in chronic pain.

This sample of 500 notes was annotated to a pseudo-ontology class hierarchy by four annotators trained in scientific writing, English, and Nursing. The hierarchy contained 13 top level classes with multiple sublevel branches, ending in 460 terminal leaf classes. Approximately 150,000 class annotations were created with approximately 136,000 linking attributes, providing approximately 80% coverage of each document. Detailed NLP procedures, including sentence identification, class assignments, scoring schemes, etc, will appear in a later publication, but a summary and focus on pain extraction is included in the next section.

Extraction of Pain Numbers. The extraction of pain numbers from the progress notes entailed initial identification of sentences that contained pain status or pain level information. The initial identification utilized the bank of class annotations and the attribute bank (class linkages). This was followed by the use of a set of regular expressions to pull only those numbers associated with pain and assign them to worst pain, least pain, today’s pain, or average pain. The bank of annotations and their class assignments were used in an exemplar or instance-based approach (implemented as a variation of past work in NLP and psychology) which compared all possible overlapping
sentence segments from the target text of size 1 to 5 words (bag of words style) to the entries in the annotation bank. The highest overlap scores (Jaccard Index: ratio of set intersection to set union) were retrieved and sorted. To capture misspellings and word roots the text instances in the annotation bank were also rendered as overlapping bi-letter sequences (spaces removed) and the n-gram word sequences in the target text were rendered similarly. The Jaccard Index overlap scores were calculated for these two sets of bi-letters. The highest score for the n-gram in the target text was retained. Class assignments for the n-gram where assigned to each word in the n-gram. Each word in the target text appeared in multiple n-grams. All the class assignments associated with a word were captured and aggregated by class. The result was a word-by-class array, for each sentence, containing values representing the class weights for each word.

Sentences containing words with non-zero scores for the classes shown in the first column of Table 1 were further processed to identify those sentences satisfying the class-attribute-class patterns shown in each line of Table 1. To satisfy a given pattern, the distances between the class, attribute, and class had to be within those seen in the annotation and attribute repository.

<table>
<thead>
<tr>
<th>Initial Class</th>
<th>Attribute</th>
<th>Target Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>has_pain_descriptor</td>
<td>pain_descriptor</td>
</tr>
<tr>
<td>Pain</td>
<td>has_pain_descriptor</td>
<td>pain_intensity</td>
</tr>
<tr>
<td>Pain</td>
<td>has_quality</td>
<td>pain_descriptor</td>
</tr>
<tr>
<td>Pain</td>
<td>has_quality</td>
<td>pain_intensity</td>
</tr>
<tr>
<td>pain_descriptor</td>
<td>has_value</td>
<td>value</td>
</tr>
<tr>
<td>pain_intensity</td>
<td>has_value</td>
<td>value</td>
</tr>
<tr>
<td>pain_descriptor</td>
<td>modifies</td>
<td>Pain</td>
</tr>
</tbody>
</table>

Once candidate sentences were identified, regular expressions were used to pull numbers from 0-10, not preceded by a letter (as c3, l4, t2, s2, representing vertebra number); Ordinal numbers (1st, 2nd, 3rd, 4th…); or number plus unit (e.g., 4 years, 4 months, 7 days, 42-year-old, 10 mg, 20 mcg, 5%, 7 percent, etc). Neighboring key words were then examined for possible assignment of numbers to worst, today, least, and average pain (e.g., “wors”, “escalat”, “most”, “today”, “now”, “least”, “lowest”, “down to”, “best”). Finally, patterns were parsed to identify such constructs as: “pain of 5 on scale of 1-10”; “pain of 10 out of 10”; “usual pain 4, low pain 1, highest pain 10 out of 10”; “low pain 5/10; high pain 10/10”; “pain 4 in range 1-10”; “pain today is 6, least it gets is 3, worst it gets is 9, on a scale of 0-10”; etc. (Additional details of these methods can be obtained from the first author.).

Recall and precision for success in selecting documents containing sentences with pain numbers were determined by blindly choosing a random sample of documents and manually determining the confusion matrix against those predicted by the algorithm.

**Beta Distribution Analysis of Pain Number Frequencies**

Patient visits containing pain numbers in both the PHA questionnaire entries and progress notes were analyzed in detail. Frequency histograms were created over the range of pain numbers extracted from progress note for each of the pain levels (0-10) from the matching PHA question. This method was applied separately for worst pain, least pain, and today’s pain for both the intake PHA (iPHA) and the follow-up visits PHA (cPHA). These histograms were often bi-modal and highly skewed in shape. To extract the two apparent modes, each histogram was estimated by the mixture of two best-fitting beta distributions. The beta distribution arises in many contexts. In order statistics, it describes the distribution of the k-th smallest variate in a random sample of n uniformly distributed variates. As such, it is applicable to finite sequences of ordered variates with bounding limits on the values of a random variate. It is only defined over the interval 0.0 to 1.0 and can be used to model distributions that are confined to an interval and exhibit left and right skewed shapes that are typically non-normal. The probability density function for the beta distribution is given by:
where $(\alpha, \beta)$ are two shape parameters and $c$ is normalizing constant. The mean and variance are $\frac{\alpha}{\alpha + \beta}$ and $\frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$, respectively. While the beta distribution is defined over the interval (0,1), we make the assumption that it can be a valid approximation for the uniformly spaced, ordered 11-point Likert pain scale used throughout the PHA questionnaire. Furthermore, we mapped the 0.0 to 1.0 beta distribution range to an extended -1 to 11 scale. This choice greatly reduced boundary effects that often reduced the quality of the fits. The fitting of the mixture of two beta distributions was performed using least squares regression implemented with the non-linear truncated Newton algorithm. For analysis of factor distributions for testing differences between patient subpopulations, maximum likelihood algorithms were used for parameter estimations, as implemented both in Stata™ using the third party module betafit, and with python scripts using the scipy.stats.fit module (http://docs.scipy.org/doc/scipy-0.17.1/reference/stats.html). These gave very similar results.

**Results**

**Extraction of pain numbers from progress notes**

NLP analysis of the 12,983 progress notes linked to intake PHAs (iPHAs) of the same day indicated that 3,891 had pain number sentences located in the text (Precision = 1.0, Recall = 0.96). Similarly, NLP analysis of the 37,348 progress notes linked to cPHAs of the same day indicated that 6,140 had pain number sentences located in the text (Precision = 0.98, Recall = 0.99). The reason not all notes had pain numbers is that pain status was consistently stated but actual pain numbers were less commonly provided. (The manner in which the progress notes where constructed was left to the physicians.) When present, they often took the form of indicating worst pain, least pain, average pain, or today’s pain, for example: “The patient indicates that pain today is 6. Least it gets is 3. Worst it gets is 9.” (See Methods for details of extraction.) Documents that contained occurrences of pain numbers averaged approximately 2 instances from these four types.

Within iPHA-linked progress notes, worst pain existed in 3,414 documents, least pain in 3,328, today’s pain in 1,542, and average pain in only 804. These were aligned with questions within the iPHA questionnaire that were also directed to these pain categories. (See Juckett et al, 2015, Table 3.) For progress notes occurring on the same day as a cPHA questionnaire, worst pain existed in 3,588 documents, least pain in 3,424, today’s pain in 3,755, and average pain in only 1,415.

**Sorting progress note pain values into PHA response groups**

Pain levels detected in progress notes were sorted into the 11 levels of pain (Likert scale) representing pain levels 0-10, as defined by the answers the patients gave on the PHA questionnaire during the same clinical appointment. These sorts were constructed for least today, worst today, current pain today, and average pain in the past four weeks. Only least pain, worst pain, and today’s pain are presented in this paper. For the least pain today question in the iPHA, the raw frequency histograms are shown in the left panels of Figure 1. In each panel, the histogram represents the pain levels extracted from the progress notes of patients that all entered a particular value for least pain in the questionnaire – as noted within the panel.

Examination of these histograms revealed the appearance of at least two distributions in each panel. One appeared to be a narrow region that was concordant with the value of the iPHA response associated with that panel. The other appeared to be a broad distribution that was offset and shifted from the first. These histograms were fit with the mixture of two beta distributions, as described in Methods. The results for iPHA least pain are shown in the right panels of Figure 1. In each panel, the two beta distributions are shown as histograms (solid bars for the narrow distributions, open bars for the broad distributions). The root mean square error is given within each panel as a percentage deviation of the fit to the original histogram. The relatively low values appear to justify the use of only two beta distributions in the fit.
Figure 1. Histograms of raw frequencies for least pain values, extracted from progress notes sorted according to iPHA-reported pain values. (Left panels) Histograms of pain values extracted from progress notes linked to PHA responses for the same visit. Each panel represents a different pain value recorded on the iPHA, as noted. Within each panel is the distribution of pain values extracted from the progress notes for those patients reporting the iPHA pain value associated with the panel. (Right panels) The two best-fit beta distributions for the corresponding panel on the left. The root mean square error (RMSE) of the fit is shown in each panel.
Table 2. Relative sizes of sync and nonsync populations and the purity (non-overlap percentage).

<table>
<thead>
<tr>
<th>Population</th>
<th>Least Pain</th>
<th>Worst Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPHA – sync</td>
<td>42%</td>
<td>36%</td>
</tr>
<tr>
<td>iPHA – nonsync</td>
<td>58%</td>
<td>64%</td>
</tr>
<tr>
<td>cPHA – sync</td>
<td>48%</td>
<td>27%</td>
</tr>
<tr>
<td>cPHA – nonsync</td>
<td>52%</td>
<td>73%</td>
</tr>
</tbody>
</table>

Figure 2. Plots of means and square roots of variance for best-fit beta distributions to histogram sets similar to Figure 1. The two components extracted for each histogram generate two vertical points at a given PHA pain level on the horizontal axis: (closed circles) peaks tracking with the PHA response; (open circles) broad peaks not tracking with PHA response. (Some distributions exceed 10 because the best-fitting beta distribution was highly skewed to the right and generated a mean above 10.) Points are plotted as pain from physician progress notes versus pain recorded on the PHA questionnaire for the same visit.  

- a) Least pain category captured in the iPHA (intake) and physician notes. The regression line for the synchronized points has slope = 0.98, $R^2 = 0.99$.  
- b) Worst pain category captured in the iPHA and physician notes. The regression line for the synchronized points has slope = 0.81, $R^2 = 0.90$.  
- c) Today’s pain captured in cPHA and physician notes. The regression line for the synchronized points has slope = 0.96, $R^2 = 0.97$.  
- d) Least pain category captured in the cPHA (cumulative) and physician notes. The regression line for the synchronized points has slope = 0.90, $R^2 = 0.96$.  
- e) Worst pain category captured in the cPHA and physician notes. The regression line for the synchronized points has slope = 0.82, $R^2 = 0.97$.  
- f) Today’s pain captured in cPHA and physician notes. The regression line for the synchronized points has slope = 0.93, $R^2 = 0.99$.  

Table 2. Relative sizes of sync and nonsync populations and the purity (non-overlap percentage).
The mean and square root of variance for each distribution in the right side panels in Figure 1 are plotted in Figure 2a. A similar compilation for worst pain from the iPHA is shown in Figure 2b and for today’s pain in Figure 2c. The least pain category (Figure 2a) reveals that the regression of the narrow peaks is linear with slope near 1.0 and has a high regression coefficient (see legend). The means of the broad distributions deviate significantly at the high PHA pain levels and they have large variances. These deviations reveal that some patients are discordant in their pain reporting. Discordancy patterns were observed in the other pain categories as well. The pattern for worst pain (Figure 2b) shows the discordancy occurs in the low pain region. The pattern for today’s pain shows discrepancies at both high and low pain, but these are not as dramatic as those for least and worst pain.

In Figure 2d-f, similar plots from analysis of the cPHA are shown. The cPHA is given approximately every six months to long-term patients. It includes questions related to outcomes and satisfaction in addition to most of the questions contained in the intake PHA.

**Selecting concordant and discordant subpopulations**

In least and worst pain categories there are regions where the two subgroups defined by the beta distribution are sufficiently separable to select two subpopulations of patients for further comparisons. That is not the case for
today’s pain category and therefore subpopulations could not be identified for this category. Those patients in the groups defining the diagonals on Figure 2 are referred to as the synchronous group (sync) because the pain levels in the progress notes are highly concordant to the values in the PHA. Those patients in groups off the diagonal are referred to as the non-synchronous group (nonsync) because their reported pain levels are discordant. Selection of subpopulations of patients from the nonsync groups can be done fairly robustly because there are regions along the horizontal axes of the histograms (see Figure 1) were the two populations do not overlap strongly. Selection of subpopulations of patients representing the sync groups yields some uncertainty because of the overlapping distribution fits. The relative proportions of patients in the sync and nonsync groups, as well as their purity (non-overlap percentage), are shown in Table 2.

**Biopsychosocial differences between sync and nonsync populations**

A question to consider is whether there are differences between the sync and nonsync subpopulations. The latent factors derived from the question items in the PHA can be used to explore differences in the biopsychosocial patterns of these populations. How patients answer the various questionnaire items indicates how severely their pain status impairs mental health, physical health, social activities, work, etc. in their daily lives. From 67 of the biopsychosocial question items in the PHA we previously identified 15 latent factors that help identify patient phenotypes. The biopsychosocial factors were all scaled between 0-10, in the same manner as pain, with low values indicating small effects of pain on a given factor and high values indicating large effects of pain on a given factor.

The differences between corresponding distribution means for the biopsychosocial latent factors of the nonsync and sync subpopulations are shown in Figure 3. The [nonsync-sync] difference values fall along a severity scale spanning positive and negative values. Positive values indicate that the nonsync subpopulation experienced more severe effects of pain on the biopsychosocial factors. Negative values indicate that the sync subpopulations experienced more severe effects of pain on the factors. The four panels represent the least and worst pain categories for both the iPHA and cPHA. Three of these pain categories have clear trends.

For the iPHA least pain category (Figure 3a), the differences for many of the factors are shifted to less impairment caused by pain, indicating that the nonsync distribution means have lower values than the sync distribution means. Many of these are significant or approach significance at the 95% confidence level, while the total trend is almost uniformly the same for all the factors. For the iPHA worst pain category (Figure 3b), the results are less dramatic, with only a few latent factors showing significant changes. Many of these exhibit a shift similar to the least pain category.

A consistent pattern appears in both the least and worst pain categories of the cPHA (Figures 3c, 3d). In the least pain category, the differences for the nonsync minus sync distribution means indicate less severe impairment for all biopsychosocial factors. Seven of the factors are significantly shifted at the 95% confidence level. In contrast, the factor differences for the worst pain category are significantly shifted to more severe impairment.

The interpretation of these findings is that the nonsync subpopulation patients in the least pain categories, who report lower pain to the physicians than they enter on the PHA (Figures 2a, 2d) are, in fact, experiencing lower severity on their biopsychosocial indicators (i.e., nonsync minus sync < 0). Similarly, the nonsync subpopulations in the worst pain category during follow-up visits are experiencing higher severity on their biopsychosocial indicators (i.e., nonsync minus sync is >0), consistent with what they report to physicians (see Figure 2e). In these three cases, the pain expressed to the physician correlates more closely to their biopsychosocial indicators of health.

**Discussion**

Patient pain reporting has been analyzed for patients seen at MPC between April, 2010 and Aug, 2014. Patients reported pain in a questionnaire (PHA) administered in the waiting room prior to their visit with the physician. Many also reported pain levels to the physician during the visit and this was often included in the physician’s progress report. Worst pain, least pain, and today’s pain values were extracted from the progress notes using natural language processing techniques and compared to corresponding questions in the PHA. Once extracted, sorted, and compared to the PHA values, two subpopulations of patients were evident; those that were concordant in their reporting on the PHA and in the physician encounter, and those that were highly discordant. These subpopulations exhibited differing biopsychosocial characteristics, as determined by their response profiles on the latent factors detected previously among the other items of the PHA questionnaire. The most important observation was that the concordant and discordant subpopulations were identified solely by the differences in their pain reporting between the questionnaire and the physician encounter. Yet, this single criterion yielded separate populations with consistently different biopsychosocial characteristics.
Although not shown, there were no significant differences in subpopulation demographic characteristics nor their ICD-9CM classifications. For the cPHA categories, there was no difference in the average number of follow-up visits among patient subpopulations. We are currently exploring the differences in treatments between subpopulations and this will be reported separately. Early indications are that drug-seeking behaviors are not involved.

It is important to recognize that the differences in outcomes and other biopsychosocial factors identified with the sync and nonsync subpopulations were derived from patient PHA reports. If one questions the veracity of patient pain reporting in the PHA, then one must be prepared to question the veracity of the other items in the PHA. The results of this study, however, reveal the complexity of the issue. If one looks at the patients that report high levels of pain in the least pain category in the PHA (high ‘x’ values on Figures 2a, 2d) and tell the physician something lower (‘y’ values), one might expect that they would also report worse outcomes and more severity in the other biopsychosocial factors in the PHA. However, these patient subpopulations report just the opposite. One would expect that if they exaggerate pain levels in the PHA, they would exaggerate other items. This does not seem to be the case. Instead, pain reporting seems to be unique and highly influenced by the differences between direct and indirect questioning. Furthermore, the fact that the biopsychosocial factors are latent factors derived from multiple items may also diminish any discordant reporting within the bulk of the PHA. Clearly, this is a matter for additional research. As one strategy to examine this phenomenon, we are currently examining outcomes and other patient histories extracted from the progress notes using NLP. It is likely, however, that case-control prospective studies will be the most valuable approach in shedding further light on this behavior.

Veracity in various settings has been the focus of several studies. Capturing patient-reported outcomes has been a focus of studies to support FDA guidance in regard to clinical trials. This issue is important to the FDA’s evaluation of drug candidates and allowable claims by the pharmaceutical industry. Several studies have also looked at subject veracity and data equivalence in other settings. A comparison between online versus in-person questioning of adolescents regarding social and inter-personal issues recently found no differences. On the other hand, a laboratory experiment examining in-person versus video conference in the context of intentional deception looked at the ability of interviewers to detect intent. The results indicated that interview modality did play a role in perception. Within the health domain, a recent study showed that some discrepancies exist among clinical data element fields of research data and clinical practice data on the same patients. The authors defined three levels of accuracy which can be useful in cohort discovery in the clinical research setting. This sampling of approaches to veracity in various settings highlights the importance of being aware of possible discrepancies in subject declarations. This is particularly important in pain reporting because decisions in pain management depend heavily on capturing patient perceptions.

The psychology and psychometrics of pain reporting has a rich literature spanning the psychosocial characteristics of a person in chronic pain to tools for patient self-reporting. To our knowledge, however, this is the first exploration of discordant pain reporting within the same clinical visit between questionnaires and face-to-face encounters. While this is an uncontrolled, retrospective study in the community setting, it has the strength of large numbers and the simultaneous capture of numerous biopsychosocial features.

**Conclusion**

Do we really know how to obtain the ground truth on a patient’s pain level? Depending on the setting, pain may be assessed differently. In the acute pain, hospital setting, pain levels are usually reported to nursing staff. In chronic pain clinics, it may be captured by physicians, physician assistants, or nurses. In observational studies, pain levels can be captured by an interviewer or by patients filling out questionnaires. In each case, the captured pain levels are typically assumed to be the ground truth; but are they? When pain is secondary to the treatment of another disease, knowing true values may prevent inducing drug dependence during the treatment of the primary illness. When chronic pain treatment is the primary medical goal, true values are critical to therapy choices. Since capturing these true values is not an obvious process, we must not be lulled into believing that entries in EHRs are the ground truth. Tools should be built with the capability of capturing the various estimates of pain perception to allow assessment and decision making, particularly in settings where pain plays a major role in healthcare.

**Acknowledgements**

We are grateful to Chris Cubbage, Cyndy Walsh of ProCare System, Inc. for their consultations and assistance in this project. We also acknowledge the annotation assistance of Alice Daily, Erin Muladore, Tracy Pierce, Alisha Williamson, and Lorrin Johnson. This work was funded, in part, by AHRQ grant R21 HS22335.
References

People, Process and Technology: Strategies for Assuring Sustainable Implementation of EMRs at Public-Sector Health Facilities in Kenya

Samuel G. Kang¹,a, MSc¹, Veronica M. Muthee, MPH¹, Nzisa Liku, MPH¹, Diana Too MPH¹, Nancy Puttkammer, PhD²

¹International Training and Education Center for Health, University of Washington, Nairobi, Kenya
²International Training and Education Center for Health, University of Washington, Seattle, WA

Key Words: EMR implementation, Clinical Decision Support, Healthcare quality

Abstract

The Ministry of Health (MoH) rollout of electronic medical record systems (EMRs) has continuously been embraced across health facilities in Kenya since 2012. This has been driven by a government led process supported by PEPFAR that recommended standardized systems for facilities. Various strategies were deployed to assure meaningful and sustainable EMRs implementation: sensitization of leadership; user training, formation of health facility-level multi-disciplinary teams; formation of county-level Technical Working Groups; data migration; routine data quality assessments; point of care adoption; successive release of software upgrades; and power provision. Successes recorded include goodwill and leadership from the county management (22 counties), growth in the number of EMR trained users (2561 health care workers), collaboration in among other things, data migration (90 health facilities completed) and establishment of county TWGs (13 TWGs). Sustenance of EMRs demand across facilities is possible through; county TWGs oversight, timely resolution of users’ issues and provision of reliable power.

Introduction

Electronic medical records (EMRs) have been embraced across public, private, and faith-based health facilities in Kenya. Perceived benefits driving uptake of EMRs in general include: optimizing documentation of patient encounters; improving communication of information to physicians; improving access to patient medical information; reducing errors; supporting administrative management; forming a data repository for research and healthcare quality improvement; and reducing use of paper¹,²,³,⁴.

The Ministry of Health (MoH) in Kenya rolled out EMRs in facilities beginning November, 2012. KenyaEMR's implementation was spearheaded by I-TECH with funding from President's Emergency Plan for AIDS Relief (PEPFAR). The system runs on an open source platform named Open Medical Record System (OpenMRS)⁵. In total, 342 instances of KenyaEMR have been implemented across four regions in Kenya: Central, Nyanza, North Rift Valley, and Western. KenyaEMR initially targeted services within the national HIV/AIDS care and treatment program, as well as related Tuberculosis (TB) and maternal child health (MCH) services. There is presently interest in expanding the system to cover other outpatient primary care services as well as inpatient services. Guidelines from the National AIDS and STIs Control Programme (NASCAP) defined the scope and workflows within KenyaEMR and all the implemented instances of the system have the same functionality. Functionality includes patient registration, triage, clinical assessment, documentation of diagnoses, prescriptions and laboratory orders, care alerts and reminders, automated monthly reporting, cohort reporting, and continuous quality indicator reporting. No interoperable provider order entry or laboratory results transfer is presently available and functionality related to billing is also absent, as services are offered at no cost to patients. The vast majority of data is captured in structured formats based upon standardized concepts⁶. Progressive software
releases have taken into consideration feedback from end users and stakeholders—including health facility staff, members of County Health Management Teams (CHMTs), and representatives of PEPFAR-supported service delivery implementing partners (SDIPs) tasked with technical assistance to public sector HIV program sites—about desired functionality.

While the anticipated benefits of adopting EMRs may be great, the processes of adopting EMRs can be highly disruptive to conventional workflows. Successfully sustaining EMRs demands a high level of organizational change centered on people, processes and technologies in healthcare, as described in various socio-technical frameworks for information systems success. This paper focuses on the specific activities undertaken by I-TECH in Kenya in support of the Ministry of Health (MoH) in Kenya focusing on the people, processes and technologies employed in EMRs implementation.

Methods

I-TECH pursued a range of strategies for assuring sustainability of EMRs implementation. We describe each strategy and its results according to the socio-technical “people, process, and technology” framework for organizational capacity to meaningfully use and sustain health information systems. The observations we report are based upon routinely-gathered monitoring and evaluation (M&E) data, as analyzed by the authors (who have been part of project implementation and M&E teams since project inception in 2012). The data has been collected through tools such as form hub, Google sheet on post implementation monitoring tool and quarterly reports shared by I-TECH regional staff “train smart” application that supports monitoring of EMRs training among other tools. Once collected, I-TECH M&E staff conducts analysis of the data on a quarterly basis for purposes of reporting on project results, successes and challenges to the funder. This study represents a synthesis of these project reports from 2012-June 2016

People:

- **Sensitization of leadership**

EMR implementation began with sensitization of local leadership on the site readiness criteria and the steps for preparing public health facilities for EMR implementation, the initial scope of the EMR functionality, the expected role of local leaders in identifying “model sites” for initial implementation (where best practices could be established) as well as subsequent “scale up sites”, and the process of engaging and communicating with target health facilities. Inadequate communication can lead to divergent understanding among health workers on the role of EMRs, and problems at the organizational level with change management. Initially, sensitization of local leadership involved outreach to provincial teams in 8 provinces; however, once Kenya adopted a new constitution in 2010, decentralizing governmental functions to 47 newly-formed County governments, sensitization was expanded to leadership in 22 of the 47 counties covering the regions of KenyaEMR implementation. Sensitization was conducted by I-TECH in collaboration with representatives from the national MoH.

- **User training**

User training is the process of building skills among target system users to use and maintain the systems. EMRs users were categorized into 3 cadres; end users; health managers; and system administrators. I-TECH supported the MoH in Kenya to develop distinct training curricula targeting each of these groups. The training was conducted in a cascaded mode. First, training of trainers (TOTs) were conducted, reaching a small pool of lead trainers in each county. Next, TOT participants helped train KenyaEMR end users as well as “KenyaEMR Champion Mentors” (referred to henceforth as ‘champions’). Champions were front-line health care workers based at the target health facilities, who were nominated based on high interest and knowledge in the EMRs, willingness to support other users in the EMRs use, and strong post-test scores during the end user trainings. Champions
represented a variety of health workers cadres (clinician, data clerk, health records officer, nurse). Champions were prepared for their role through supplemental mentorship trainings. End user trainings also targeted diverse cadres involved in HIV care and treatment programs during training of 3-5 days. The management at the facility was targeted through the health managers’ curriculum which equipped them to provide oversight for the use of the EMRs at the facility level. The system administrators training was designed to prepare IT specialists based within County Health Management Teams (CHMTs) and large health facilities to trouble shoot EMRs hardware and local network problems, data backup and restoration procedures, management of user accounts, and other technical and “help desk” issues.

- **Multi-disciplinary teams**

Health facilities have multi-disciplinary teams (MDTs), supported by the PEPFAR SDIPs to address cross-cutting issues in HIV/AIDS service delivery. The MDTs comprise the doctors, clinicians, nurses, peer educators, mentor mothers, nutritionists and counselors working within the Comprehensive Care Center (CCC), the department specializing in HIV/AIDS care. Each MDT is chaired by the doctor or other clinician who directs the CCC and other services (the “facility in-charge”). EMR champions and facility in-charges facilitated the inclusion of the EMR-related topics alongside other agendas discussed by the MDTs. Examples of EMR-related issues addressed by MDTs include: roles and responsibilities in oversight of EMR activities, planning for legacy data migration, routine data quality assessment, and planning to reduce various barriers to EMR use.

- **County Technical Working Groups**

As part of the transition of EMRs to sustained local leadership and ownership, I-TECH collaborated with the SDIPs and CHMTs to form eHealth Technical Working Groups (TWGs) to oversee the EMRs agenda within each county. The TWGs deliberately comprised officials with authority and capacity to undertake timely decision making. Each county TWG is typically chaired by the County Director of Health and in his or her absence, by the County Health Records Information Officer (CHRIO). Members include other representatives from the County Health Management Team (CHMT), from EMR technical assistance partners, and from PEPFAR SDIPs.

Given the access of the TWGs to the county leadership, the groups have facilitated adoption of a number of proposals. Examples include recommendations that senior county leadership take into account EMR skills and experience when managing health worker employee transfers between health facilities (i.e. transferring EMR-experienced health workers to other sites with EMRs), that they plan for EMR-related costs when determining budget allocations during the annual work planning meetings. Allocated budgets should support EMRs maintenance, EMRs review meetings, and routine data quality assessments, among other EMR-related activities.

- **OpenMRS community building**

In an effort to increase the local support available for further KenyaEMR software enhancement and support I-TECH Kenya convened a meeting of OpenMRS implementers in Kenya and organized OpenMRS and KenyaEMR boot camps. The implementers meeting targeted developers and implementers from organizations customizing EMRs on the OpenMRS platform, in order to share best practices and identify areas of collaboration. The boot camps targeted university students pursuing information technology and computer science courses, in their final or penultimate year of study. This was to sensitize and orient them to the basics of the OpenMRS platform and to stimulate interest in health informatics careers.
Process:

- **Data migration**

  Public sector HIV/AIDS care and treatment programs in Kenya date back to 2005. Facilities offering such care have patient records since the inception of the service by the MoH. The current policy on paper records requires their preservation up to a period of 10 years. As facilities adopted EMRs, there was a need to have the historical patient data captured on the EMRs, to support continuity of care. Effective HIV/AIDS care can only occur when clinicians have access to key historical data for each patient and can track their response to treatment over time.

  To understand the effort needed in legacy data migration, we sampled the largest of the KenyaEMR databases from a large referral hospital in the Nyanza region of Kenya, with 26,971 patients’ records as of January, 2016. The total number of patient visits from the instance was 612,521 and the total number of observations (i.e. individual data elements) was 6,905,588. Therefore, the average number of visits per patient was 22 (612,521/26,971), and the average number of data elements captured per visit was 11 (6,905,588/612,521). With exception of the patients’ demographic information which is mostly captured once, legacy data migration would need to capture 242 data elements per patient. The number of patients per facility ranges between 500 and 26,000 patients. Together, this implies a heavy load of legacy data migration across KenyaEMR sites.

  To support the transfer of paper records to EMRs, data clerks were recruited and orientated to the data capture process to ensure that paper-based records were migrated into the electronic record. Data migration efforts were initiated immediately after the end user training. I-TECH collaborated with CHMTs and SDIPs to conduct intensive legacy data migration initiatives over a period of 3 months, whereby supplemental data clerks were recruited and assigned to specific health facilities.

- **Routine Data Quality Assessment (RDQA)**

  RDQA refers to the continuous assessment of the quality of data contained in the EMRs. It involves comparing patient data captured on the EMRs against the data contained in the source paper records for consistency of information. KenyaEMR provides an automated RDQA report that draws a random sample of patient records from the EMR, to enable comparison with other data sources, with a sample size intended to estimate the concordance level of data within +/-5%. Using the RDQA report, facility personnel or external assessors can compile comparison data values within an Excel-based workbook which is pre-programmed with charts and graphs to show RDQA results. Upon completion of each RDQA, sites identify action plans to address data quality gaps. Sites with weak concordance of data (<80%) are encouraged to follow standard operating procedures for data review and cleaning, and to undertake repeated RDQAs every 3-4 months until data quality reaches expected levels.

- **Point of Care (PoC) adoption**

  EMRs implementation in Kenya occurred in either PoC or Retrospective Data Entry (RDE) mode. PoC refers to a setup where a computer terminal is implemented at every patient service point within CCC, TB and MCH services. RDE refers to a setup where clinicians use paper records to serve patients and later have the paper records captured into the EMR at a single centralized data entry point by themselves or designated health facility workers. 

  PoC adoption permits use of clinical decision support functions within KenyaEMR. Available alerts include: patient due for viral load testing; patient eligible for ART; patient lost to follow-up. These features promote timely action towards healthcare quality improvement.
Upon initial assessment for EMR implementation readiness, facilities that had physical security gaps (e.g., no grills on windows, no bugle-proof doors) were recommended for RDE implementation. County and facility management were informed of the gaps. At sites where the gaps were remediated, subsequent transition from RDE to PoC implementation was possible.

**Technology:**

- **Successive releases**

  The initial KenyaEMR software release was based on an HIV/AIDS module for use in CCCs. This was augmented over time in response to the MoH users’ needs and with support from I-TECH to include maternal child health (MCH) care and Tuberculosis (TB) care modules. Additional features include Data Quality features, Reports Generation among others. There is presently interest in expanding the system to cover other outpatient primary care services as well as inpatient services.

- **Power sources**

  KenyaEMR implementation involved readiness assessments at each health facility to identify presence or absence of resources needed for EMR implementation. The assessments addressed physical security, electricity, presence of existing EMRs, and other issues. Where facilities were estimated to have electricity up to 75% of the time and met other required factors, the site was approved for implementation. However, power availability was often over-estimated. In reality, power supply is highly dynamic, with frequent power blackouts arising from normal rationing, other faults, or power cuts due to non-payment of bills.

  I-TECH collaborated with the CHMTs and SDIPs to analyze and compare costs of alternative power sources. Kenya has sunlight for more than 90% of the day with exception of the wet season. In the wet season, rationing of power is less common due to sufficient supply of water to the hydroelectric dams that provide electricity. Solar power was found to be cheaper than setting up a generator (US$7,940 vs US$19,250 for a solution supporting a load of up to 1.5 Kilowatts, or US$18,940 vs. US$74,690 for a solution supporting a load of up to 6 Kilowatts, each running for 8 hours per day over a period of 10 years (Table 2). Having alternative backup sources of power at health facilities has been identified as a critical factor preventing the piling up of backlogged patient records.

**Table 2:** Comparative costs for setting up and running backup source of power (8 hours per day x 10 years)

<table>
<thead>
<tr>
<th>KW</th>
<th>Solar (cost in USS)</th>
<th>Generator (Cost in USS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.5</td>
<td>6</td>
</tr>
<tr>
<td>Fuel Cost per KWh</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fuel Cost per year</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fuel Cost in 10 years</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Number of services (@200hrs)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Service Cost</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Battery Replacement</td>
<td>1,560</td>
<td>2,340</td>
</tr>
<tr>
<td>Total Running Costs</td>
<td>1,560</td>
<td>2,340</td>
</tr>
<tr>
<td>Installation Cost</td>
<td>6,380</td>
<td>16,600</td>
</tr>
<tr>
<td>Total Cost of Ownership</td>
<td>7,940</td>
<td>18,940</td>
</tr>
</tbody>
</table>

681
Results

People
I-TECH supported sensitization of 678 participants from the county leadership. This bore fruits in their facilitation of EMR related activities, including: recommendation of health workers to participate in planned EMR training; supervision of EMR use at the facility level; convening review meeting across the regions where facilities shared their experiences and best practices.

I-TECH further supported the training of over 2561 health care workers (42 doctors, 189 lab technologists, 868 clinical officers, 777 nurses, 370 health records and information officers and 315 other cadres) in 297 trainings across 22 counties in the four regions of Kenya (Table 3).

Table 3. Distribution of trained health workers for EMRs use across North Rift valley, Nyanza, Western and Central regions of Kenya

<table>
<thead>
<tr>
<th>Health Care Workers Cadres</th>
<th>Numbers trained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctors</td>
<td>42 (2%)</td>
</tr>
<tr>
<td>Laboratory Technicians</td>
<td>189 (7%)</td>
</tr>
<tr>
<td>Clinical Officers</td>
<td>868 (34%)</td>
</tr>
<tr>
<td>Nurses</td>
<td>777 (30%)</td>
</tr>
<tr>
<td>HRIO</td>
<td>370 (14%)</td>
</tr>
<tr>
<td>Other</td>
<td>315 (12%)</td>
</tr>
</tbody>
</table>

Champions conducted one-on-one sessions with users to train them on specific tasks or to bring newly appointed staff members up to speed on the EMR. Altogether, Champions reported mentoring 1754 facility staff mentored across 180 health facilities. Staff also interacted with the KenyaEMR “help menu” containing job aids, video tutorials, and two self-paced eLearning modules. In addition, Champions and KenyaEMR end users initiated region and county specific user support groups, via the “Whatsapp” mobile phone application

The “Whatsapp” groups have been formed across all 22 counties, and serve as a forum for peer-to-peer trouble shooting and bug reporting. On average, about 10 issues are resolved per county per month across the 22 counties implementing KenyaEMR. Facility staffs are further empowered through services such as toll free lines through which they can receive technical support and an online help desk management and bug reporting tool called Redmine.

Process
Out of the 342 implemented KenyaEMR instances, 90 health facilities had achieved 100% data migration while 252 were yet to fully migrate the legacy data into the EMR, by January 2016. The persisting need for legacy data migration has been occasioned by factors such as: downtime due to fluctuation of electricity; lost capacity following transfer of EMR trained health workers; increased workload due to double entry of patient records in both paper-based and EMR data systems; lean staff across health facilities; and slow data migration of patient records contained in non-standard MoH tools.

Specific data cleaning efforts have since been initiated to address identified gaps. In an analysis of baseline vs. follow up RDQAs conducted across 27 health facilities, there were 2,411 patient charts reviewed during baseline RDQAs and 2,381 patient charts reviewed during follow up RDQAs. At baseline, the concordance of data for mandatory data elements was lowest for dates of first and last CD4 test (13% and 11%, respectively) and highest for program of enrollment and patient sex (79% and 84% respectively). Follow up RDQA scores improved on average by 1.8 points on a 20-point scale across the 27 health facilities.

Technology
Overall, 208 (61%) initial KenyaEMR implementations were of the PoC model while 134 (39%) were of the RDE model. Additional 27 facilities have been assessed for upgrade to PoC implementation. This has been
facilitated by continuous improvement of physical security paving way for deployment of additional computing terminals.

Although power supply is a principal challenge to EMR use, only 50 out of 342 health facilities have an alternative power source that is maintained by the health facility (8 in the Central region, 16 in the North Rift Valley region, 9 in the Nyanza region, and 17 facilities in the Western region). This has greatly facilitated the continuity of EMR use in the absence of power from the national grid.

Overall results
An impact evaluation of the KenyaEMR implementation in improving quality of HIV/AIDS care is planned. Presently, qualitative feedback from EMRs users shared during county-level TWG and EMR review meetings indicate general satisfaction with KenyaEMR. Champions and end users have noted, for example: improved ease of retrieving patient records, reduced data entry errors owing to the presence of validation rules at data entry, reduced duplication of patient records due to easy-to-use patient lookup feature, increased ability to easily detect and resolve duplicate records based upon a feature that identifies potential duplicate records using patient demographic data.

Discussion

Lessons learned for EMR success
Our experience with EMR implementation at 342 health facilities in Kenya underscored many lessons learned for sustainable success of the systems. The process of EMR implementation brings with it numerous expectations, some of which are realistic and others which are not, such as: ability to immediately eradicate use of paper records\(^{13}\), automatic reduction of errors\(^{14}\), patient data security and confidentiality, and even loss of jobs for health workers. There is need to manage the expectations and demystify myths well in advance to reduce reluctance to use EMRs, calm users’ frustrations, and set reasonable expectations about the effort required for successful EMR implementation. Further, specific efforts should be made to actively monitor the successes and challenges encountered during the implementation stage. Health workers expectation has been cited as one of the factors that affect users’ satisfaction in EMRs\(^{15, 16}\). Others include practicality of use, impact on delivery of care, organizational support and variation from familiar practice\(^{17}\).

Without buy-in from the senior leadership it becomes a struggle getting health workers to use EMRs and an EMR project can easily be labeled as a failure\(^ {18, 19, 20, 21}\). Health facilities have different dynamics in uptake of EMRs and it is important to actively engage health facility leadership before proposing implementation plans\(^ {22}\). The process of EMRs implementation is cost intensive in the hardware acquisition, data migration, users training, provision of alternative power sources, among other key activities\(^ {23}\). A costing evaluation covering I-TECH’s direct costs of EMR implementation showed an average site-specific cost of US$9,879 per facility to implement an EMR\(^ {24}\). It is paramount that county teams and facilities identify sponsors beforehand to ensure that the funds available can support the different EMRs projects to completion. Recurrent costs including replacement of hardware should further be considered and planned for by the project owners. Besides sourcing funds, the benefits of such interventions should be clear to win the support of the sponsors.

Creating avenues for communication between the system users and the support and development teams ensures that the users never feel stranded and without help. In addition it creates avenues for the EMRs to improve with the growing end-user needs. This further builds confidence for the EMRs leading to indirect influence on the acceptance and readiness to use EMRs. Past experiences show that EMRs have failed where user acceptance needs have not been addressed\(^ {25}\). Our experience showed a strong enthusiasm for forums where peers can engage and discuss issues affecting their EMR use. These have reduced the time taken to resolve issues, and have led to knowledge sharing and better capacities across health facilities.

Future Activities
Legacy data migration has proven to be a formidable challenge, but steady progress has been made. As of June 2016, I-TECH, NASCOP, CHMTs and SDIPs were engaged in an intensive push to complete legacy data
migration at all sites with remaining backlogs of historical records. Completion of data migration will allow health facilities to engage in data quality improvement activities, and to facilitate transition to Point of Care (PoC) system use which takes advantage of the clinical decision support function within KenyaEMR.

Several SDIPs which have experienced benefits from the EMRs implementation are in advanced stages of planning to support alternative power sources. For example, one partner plans to install solar generated power and voltage stabilizers across 13 health facilities with a high number of HIV/AIDS patients, to maximize system uptime and minimize risks to hardware and data due to power surges. Efforts are underway to move towards paperless operations at sites which have demonstrated strong EMR data quality and with stable power availability. For example, one partner plans to pilot a reduced use of paper tools for parallel data entry in 21 health facilities, an important step in fulfilling the promise that EMRs can enhance rather than hinder the productivity of health care workers.

Users’ needs have further grown leading to extension of the KenyaEMR roadmap to include additional functionality including; support for outpatient and inpatient departments, data exchange between different systems; EMRs to laboratory information systems, EMRs to pharmacy information systems, and incorporation of national unique patient index (NUPI) that would facilitate patient tracking and identification across health facilities.

**Conclusion**

EMR uptake has grown over time in Kenya and the expectations from users have grown in equal measures. Timelines in EMRs upgrade to include additional features being requested by users will attract confidence and drive data use to greater heights. Availability of reliable power from the main grid and alternative sources will ensure sustained uptime of the EMRs and reduce the temptation to fall back to papers. Data exchange between EMRs will further eliminate duplication of effort in patients’ registration, further strengthening the users’ experience. Overall, the change process in EMR implementation ought to respond to the people, process and technology issues to increase the chances of successful EMR implementation.

**References**

11. Church K, de Oliveira R. What’s up with whatsapp?: comparing mobile instant messaging behaviors with traditional SMS. In Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services 2013 (pp. 352-361). ACM.
Long-Term Impact of an Electronic Health Record-Enabled, Team-Based, and Scalable Population Health Strategy Based on the Chronic Care Model

Kensaku Kawamoto, MD, PhD, MHS1, Kevin J Anstrom, PhD2, John B Anderson, MD, MPH3, Hayden B Bosworth, PhD4,5, David F Lobach, MD, PhD, MS3,4, Carrie McAdam-Marx, MSCI, PhD, RPh,6 Jeffrey M Ferranti, MD, MS7,8, Howard Shang8, Kimberly S H Yarnall, MD3

Departments of 1Biomedical Informatics and 6Pharmacotherapy, University of Utah, Salt Lake City, Utah; Departments of 2Biostatistics and Bioinformatics, 3Community and Family Medicine, 4Medicine, 5Psychiatry and Behavioral Sciences, and 7Pediatrics; and 8Duke Health Technology Solutions, Duke University, Durham, NC

Abstract

The Chronic Care Model (CCM) is a promising framework for improving population health, but little is known regarding the long-term impact of scalable, informatics-enabled interventions based on this model. To address this challenge, this study evaluated the long-term impact of implementing a scalable, electronic health record (EHR)-enabled, and CCM-based population health program to replace a labor-intensive legacy program in 18 primary care practices. Interventions included point-of-care decision support, quality reporting, team-based care, patient engagement, and provider education. Among 6,768 patients with diabetes receiving care over 4 years, hemoglobin A1c levels remained stable during the 2-year pre-intervention and post-intervention periods (0.03% and 0% increases, respectively), compared to a 0.42% increase expected based on A1c progression observed in the United Kingdom Prospective Diabetes Study long-term outcomes cohort. The results indicate that an EHR-enabled, team-based, and scalable population health strategy based on the CCM may be effective and efficient for managing population health.

Introduction

Chronic diseases such as diabetes mellitus are the leading cause of death in the United States,1 with almost half of all adults having at least one chronic illness.2 Therefore, there is an urgent need to identify and deploy effective and scalable population health strategies for chronic disease. The Chronic Care Model (CCM) is a promising approach for population health management that encompasses six elements designed to make the care of chronic conditions proactive, planned, and population-based rather than acute and reactive: delivery system design, clinical information systems, decision support, organizational support, self-management support, and community interaction.3,4

The CCM has been generally shown to be effective for improving process measures in published studies, but the CCM has not been found to be uniformly effective for improving clinical outcome measures.5 Similar promising, but non-uniform, results have been reported when various components of the CCM were implemented in various combinations.5 Moreover, under traditional fee-for-service payment models, the financial benefits of improved population health can accrue to insurers, while the costs of implementing improved population health management generally fall on providers.5 Therefore, cost-effective, scalable approaches are needed that do not require significant additional resources. For example, while Piatt et al. demonstrated that a CCM-based intervention can significantly improve process and outcome measures for diabetes management, this intervention required resources that would not be generally available to primary care practices, such as chart auditors to provide performance feedback reports and on-site diabetes educators supported by the investigators.6 Moreover, the increasing shift from “volume-based” to “value-based” payment models7 is making it imperative for health care providers to efficiently improve the health of their patient populations. To succeed in this new environment, health care providers must develop population health strategies that deliver greater value through improved outcomes and lower costs.

Given the promising, but non-uniform, results from previous evaluations of the CCM, as well as the need for further evaluations of pragmatic, scalable population health strategies based on this model, the present study was conducted. We evaluated a CCM-based population health program that was in operational clinical use at 18 primary care practices of the Duke University Health System from January 2009 until June 2011, when a new electronic health record (EHR) system was implemented. While the population health program encompassed several chronic
conditions as well as preventive care, this study was focused on diabetes due to its inclusion in the population health program from the initial program deployment and the availability of a widely accepted physiological measure of disease control in the form of hemoglobin A1c (A1C) levels. A1C levels reflect average blood glucose levels over a three month period and are strongly correlated with adverse outcomes.

Because diabetes is a progressive condition characterized by worsening insulin resistance, A1C levels typically increase over time despite treatment. For example, a literature review of the expected progression of type 2 diabetes found that A1C levels increase by an average of as much as 0.5% per year (absolute change) even with therapy. While such disease progression is managed through therapy intensification, including the addition of new therapeutic classes such as insulin, A1C levels are still generally expected to increase over time. For example, based on long-term data from the United Kingdom Prospective Diabetes Study (UKPDS), and as incorporated into the UKPDS Outcomes Model, an individual starting with an A1C of 7% is expected to experience an A1C increase of approximately 1.5% per year (relative change) despite therapy. Thus, while reducing A1C levels (e.g., to < 8% or < 7%) is an important clinical goal, maintaining or slowing the increase in A1C levels is also an important goal.

Prior to the introduction of the EHR-enabled, team-based, and scalable population health strategy in this study, a labor-intensive, disease registry-based approach had been in place for diabetes management at the study sites. The objective of this 4-year longitudinal study was to evaluate whether the more efficient and EHR-based population health strategy could maintain or potentially surpass the glycemic control achieved under the previous, labor-intensive approach. A before-after study design was used in conjunction with an interrupted time series analysis.

Methods

**Study Site and Participants.** The study was conducted at 18 primary care practices of Duke Primary Care (DPC), which is a community-based primary care system affiliated with the Duke University Health System and centered in Durham, North Carolina. DPC serves a seven-county area with internal medicine, family medicine, pediatrics, and urgent care services. DPC is comparable to a community primary care system, with no residents and some teaching provided for medical students, physician assistant students, and nurse practitioner students. During the intervention period, DPC was staffed by over 100 physicians and 15-20 physician assistants and nurse practitioners.

Study participants were adult patients with diabetes mellitus types I and II who were patients of DPC during the study period (2007-2010). Detailed inclusion criteria are as follows: (i) at least 18 years old as of January 1, 2007; (ii) had at least one ICD9CM diagnosis for diabetes as defined by the National Quality Forum in the first historical control year (2007) and also at least one diabetes diagnosis in the year prior (2006); and (iii) were seen at least once in the study practices in each of the four study years, as well as in the year prior (2006). The ICD9CM diagnosis codes used to identify diabetes were 250, 250.X, 250.XX, 357.2, 362.0, 362.0X, 366.41, 648.0, 648.00, and 648.0X. The study was approved by the Duke University Institutional Review Board (protocol # 00042607).

**Study Period.** The study period was January 2007 through December 2010, with the intervention rolled out across the study sites in January 2009. The study period was ended in 2010 because the EHR system was changed in 2011.

**Interventions.** The intervention was a Chronic Care Model (CCM)-based population health program for diabetes. This multi-faceted intervention included the use of an EHR system for pre-visit planning, the implementation of a point-of-care clinical decision support (CDS) system, the implementation of a quality reporting system, team-based care, active engagement of patients, and provider education. Care team development and patient engagement were integral to the intervention. The care team consisted of clinicians (physicians, physician assistants, and nurse practitioners), as well as certified medical assistants, licensed practical nurses, and registered nurses. Patient self-management was supported through pre-visit planning and education from clinical support staff. The population health program was piloted on a small scale at the end of 2008 and rolled out across all sites in January 2009.

In order to promote external generalizability and scalability, the population health program was designed and implemented in a manner that could potentially be replicated widely. In particular, the need for labor-intensive tasks such as manual data entry was minimized, so that the approach could be sustained in typical clinical settings with existing clinical staff. Moreover, the point-of-care CDS system and quality reporting system were implemented using a software module that can be integrated with various EHR systems through a standard Health Level 7 (HL7) Web service interface. Consequently, the informatics approaches leveraged in this study could potentially be deployed to other institutions and information system environments.
The details of the CDS system have been previously described.\textsuperscript{18} Briefly, the CDS system was available as a link within the EHR system and provided a dashboard for managing diabetes, including care metrics (e.g., appropriate laboratory monitoring), the patient’s current status on the metrics (e.g., not due, due now, almost due), relevant data (e.g., date and value of most recent laboratory test), and a summary of relevant guideline recommendations (Figure 1). This CDS system initially supported diabetes, hypertension, and health maintenance, and it was later enhanced to support other conditions including congestive heart failure, asthma, and chronic kidney disease. Diabetes care recommendations were based on American Diabetes Association guidelines.\textsuperscript{12} As noted earlier, this study was focused on diabetes \textit{a priori} due to its inclusion in the population health program from initial deployment and the availability of a widely accepted physiological measure of disease control (A1C levels).

![Figure 1. Screenshot of point-of-care CDS system.](https://clinapp6.duhs.duke.edu:8081)

CDS recommendations could be obtained in a disease-specific manner (e.g., using the “Diabetes” tab in Figure 1), or for all supported conditions at once by selecting the “All” tab. This point-of-care CDS system was designed to align with the findings from a previous systematic review of CDS interventions, which found that computer-based CDS interventions that provided actionable recommendations automatically and at the appropriate point in clinical workflow significantly improved clinical practice in over 90\% of randomized controlled trials.\textsuperscript{19}

In a typical clinic visit, clinical staff used the CDS system to generate the care reminders, which were printed and provided to clinicians for use during the encounters. This approach worked well due to the routine use of a visit billing form to which the printed reminders were attached. Alternatively, clinicians retrieved the care reminders directly through the EHR. Furthermore, standing orders were in place for clinic staff to perform required interventions per protocol and as identified by these tools. For example, if the CDS system identified that a patient was in need of an A1C test, nurses were empowered to obtain the required test. This team-based care involved the use of existing clinical staff (i.e., clinicians, nurses, and medical assistants), rather than the addition of care providers who would generally not be available in a typical primary care practice (e.g., dedicated dietician, pharmacist, etc.). Most data points (e.g., laboratory test results, blood pressure values, vaccinations, eye exams) were retrieved from the EHR and did not require additional data entry. Clinic staff facilitated any supplemental data updates (e.g., to document that an eye exam was not indicated because of blindness), and the clinic staff also facilitated the completion of needed care interventions. Clinicians engaged patients through discussion of the care reminder summary and needed interventions.

The quality reporting system (Figure 2) provided the clinic managers and designated nurses of each clinic with monthly reports on performance measures, which were then used to recall patients to obtain necessary care.
The goal of this intervention component was to identify patients who were outliers in relevant physiological measures (e.g., A1C levels), so that designated nurses could increase patient engagement in their care, identify and seek to address barriers, and schedule follow-up appointments. The quality reporting system was included in the intervention based on a previous randomized controlled trial that showed that adding a clinician feedback component increased the effectiveness of a CDS system. Implementation of this approach was challenging, however, due to the labor-intensive nature of the required follow-up and difficulty in reaching some patients.

Figure 2. Quality reporting system with views for clinic group (top), clinic (middle), and clinician (bottom).
Pre-Intervention Population Health Management Approach. During the historical control period, usual care was provided using a labor-intensive approach centered around the use of a commercial disease registry tool known as DocSite. In our implementation, DocSite was a stand-alone disease registry separate from the EHR, with data imports available only for demographic data. Clinical staff were required to manually enter all other data (e.g., blood pressure values, immunization records, and laboratory values). In a typical clinical visit, a visit planner similar to that in Figure 1 was printed, which included information on what was due and what was done previously. A medical assistant would fill out information on needed data points (e.g., relevant laboratory results), and clinical staff would transcribe that data into the registry. The main differences of the study intervention compared to the historical control were (i) little or no manual data entry was required for the intervention, versus significant manual data entry required for the historical approach; (ii) integration of the population health management system with the EHR; and (iii) the availability of more sophisticated algorithms in some cases, for example for vaccination logic.

From an end-user perspective, the functionality of the DocSite system was quite similar to the study intervention except for the extensive need for manual data entry. Specifically, the DocSite system provided clinicians with diabetes care recommendations during clinic visits and outlier reports were available for identifying patients in need of targeted follow-up. Thus, even a demonstration of non-inferiority would show value, as the intervention significantly reduced the labor resources required for managing the health of patients with diabetes.

Evaluation Measures. The primary evaluation measure was the change in A1C levels among study participants during the intervention period compared to the historical control period. A potentially important confounder in the use of clinically measured A1C levels is that diabetes care guidelines recommend that A1C levels be measured more frequently among individuals with higher A1C levels. Pursuant to these guidelines, for example, the point-of-care CDS system reminded clinicians to test for A1C levels every 3 months if the last A1C level was >= 7%, and only every 6 months if the last level was < 7% (Figure 1). In order to address this potential source of confounding, as well as to reduce the impact of any random measurement errors, we estimated the A1C level of each patient on the 15th day of each month. If the patient had a test on the 15th, that value was used. Otherwise, the value was estimated through the interpolation of the nearest values (i.e., the latest test result from before the 15th and the earliest test result from after that date). This approach ensured that all patients had an equal number of A1C level estimates (48 values, one for each month of the four year study period).

A secondary study measure was the proportion of encounters at study clinics by study participants which resulted in use of the CDS system for diabetes. The system was considered to have been used for the encounter if system use logs indicated that the diabetes module of the CDS system had been activated for the patient on the day of the encounter or up to three days prior to the encounter in preparation for the visit.

Estimate of Expected A1C Progression. We estimated the expected progression of average A1C levels in the study population over the study timeframe using an A1C progression equation in the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model© (https://www.dtu.ox.ac.uk/outcomesmodel/). This equation was developed based on the UKPDS long term outcomes data, with patient level A1C data from our study cohort used in the model as the input for time zero A1C. We did not use this estimate for hypothesis-testing purposes, but we used it to contextualize our findings in the absence of a comparison group.

Analysis Approach. The study cohort was identified using a guided database query tool known as DEDUCE, which provides a Web-based interface for identifying patient cohorts and for obtaining approved data from the Duke University Health System’s data warehouse. Most study data were obtained through this tool. Data on the utilization of the point-of-care CDS system was obtained from the Duke Clinical Data Repository.

A software program was developed in Java to prepare the data in the format required for statistical analyses. Statistical analyses were conducted using SAS Enterprise Guide Version 4.3 (SAS Institute, Inc., Cary, NC, USA). Usage rates for the point-of-care CDS system were summarized graphically and using descriptive statistics. To estimate the changes in A1C levels during the pre- and post-intervention periods, a mixed linear model analysis was performed using the PROC MIXED procedure with maximum likelihood estimation. In the analysis, the patient was the class, the number of years since the beginning of the evaluation period was the predictor, and the A1C level was the dependent variable. To evaluate whether the changes in A1C levels in the intervention and pre-intervention periods were different, the PROC MIXED procedure was used with maximum likelihood estimation, with the patient as the class and with the years before or after January 2009 as the predictors. For all analyses, indicator variables for the months of the year were included in the models to adjust for seasonal variation, and sensitivity...
analyses were conducted omitting these variables. To the best of the investigators’ knowledge, no other population health management or clinical practice changes believed to have influenced diabetes care and outcomes were undertaken during the study timeframe.

The sample was restricted by the number of patients meeting inclusion criteria seen at the 18 primary care clinics. With a sample of 6,768 subjects, for the primary measure, we estimated a power of >99% to detect an improvement in A1C levels during the intervention period at least 0.5% greater than the change during the control period. These calculations were based on a one sample t-test with a 2-sided type I error rate of 0.05.

Results

There were 6,768 individuals with diabetes mellitus who received care at the study sites, met the study inclusion criteria, and were included in the analysis. All patients meeting inclusion criteria were included. These individuals had an average of 8.7 distinct A1C test results on record during the four year study period. The average A1C level for this population during the study period was 7.2%, with a standard deviation of 1.4%.

Usage of the Point-of-Care CDS System. Figure 3 shows the usage of the point-of-care CDS system for the study population for diabetes care. As noted, prior to the intervention start date (January 2009), there was limited usage of the CDS system for diabetes care for several months due to pilot use of the system by some sites. To roll out the population health intervention, the Duke Primary Care leadership notified each clinic of the switch-over from the manual data entry-driven approach to the more automated intervention approach, and basic on-site education and training were provided at each practice by clinic staff. Immediately following the start of the intervention period, usage rates increased to approximately 45%, and steadily increased to approximately 70% by the end of the study. Stratified at the clinic level, the highest usage rate in the final month of the study was 88% at one clinic; 5 clinics had usage rates > 80%; 6 clinics had usage rates > 70% and < 80%; 3 clinics had usage rates > 60% and < 70%; and 4 clinics had usage rates < 60%. Usage therefore was high but not universal. During the intervention, usage of the tool was not monitored at an individual provider level. While we did not measure the time required to use the system during this intervention, there was significant manual work that was required for data entry in the pre-intervention period. Based on expert opinion (JBA, Duke Primary Care Chief Medical Officer), we estimate that in the post-intervention period there was a savings of at least an hour a day for the clinical support staff at each clinic, depending on the size of the practice and number of providers.

Figure 3. % of encounters at study sites by study participants with CDS system use for diabetes care. Month 25 = January 2009.
**Hemoglobin A1c Levels.** Figure 4 provides an overview of the study population’s A1C levels during the historical control and intervention periods. The A1C levels remained relatively unchanged during both the control and intervention periods. During the pre-intervention period, the A1C started at 7.17% in January 2007 and ended at 7.20% in December 2008. During the intervention period, the average A1C level remained essentially unchanged, starting and ending at 7.24%. Included in Figure 4 is the expected A1C progression from the UKPDS long term outcomes study, with an expected A1C level of 7.59% (0.42% absolute increase) by December 2010.

![Figure 4. Average A1C levels over time for study population (2007-2010)](image)

Table 1 provides the model for the A1C level during the pre-intervention period, and Table 2 provides the model for the A1C level during the intervention period. As noted, the slope for the A1C level was 0.01%/year for the pre-intervention period and -0.02%/year for the intervention period. This difference in slope was statistically significant (p = 0.01). However, we do not consider this difference in slope to be clinically significant.

The intercept was higher in the intervention period compared to the pre-intervention period. However, the overall trend in A1C was essentially flat throughout the study period, starting at 7.17% in January 2007, going to 7.24% in January 2009 (the first month of the intervention), and ending at 7.24% in December 2010.

Sensitivity analyses were also conducted in which indicator variables for the months of the year were removed. These sensitivity analyses without adjustment for seasonal variation resulted in similar findings as the primary analyses with seasonality adjustment.
### Table 1. Model for A1C level during pre-intervention period (2007-2008)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.1494</td>
<td>0.01773</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>yearSinceJan2007</td>
<td>0.005152</td>
<td>0.003045</td>
<td>0.0907</td>
</tr>
<tr>
<td>Jan</td>
<td>-0.00323</td>
<td>0.007964</td>
<td>0.6855</td>
</tr>
<tr>
<td>Feb</td>
<td>0.009910</td>
<td>0.007879</td>
<td>0.2084</td>
</tr>
<tr>
<td>Mar</td>
<td>0.01454</td>
<td>0.007801</td>
<td>0.0624</td>
</tr>
<tr>
<td>Apr</td>
<td>0.01267</td>
<td>0.007730</td>
<td>0.1013</td>
</tr>
<tr>
<td>May</td>
<td>-0.00084</td>
<td>0.007668</td>
<td>0.9131</td>
</tr>
<tr>
<td>Jun</td>
<td>-0.01552</td>
<td>0.007613</td>
<td>0.0415</td>
</tr>
<tr>
<td>Jul</td>
<td>-0.02936</td>
<td>0.007566</td>
<td>0.0001</td>
</tr>
<tr>
<td>Aug</td>
<td>-0.03907</td>
<td>0.007528</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Sep</td>
<td>-0.04232</td>
<td>0.007498</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Oct</td>
<td>-0.03694</td>
<td>0.007476</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Nov</td>
<td>-0.02196</td>
<td>0.007463</td>
<td>0.0033</td>
</tr>
<tr>
<td>Dec</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

### Table 2. Model for A1C level during intervention period (2009-2010)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.2869</td>
<td>0.01703</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>yearSinceJan2009</td>
<td>-0.01916</td>
<td>0.002727</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Jan</td>
<td>-0.01099</td>
<td>0.007131</td>
<td>0.1232</td>
</tr>
<tr>
<td>Feb</td>
<td>0.01128</td>
<td>0.007055</td>
<td>0.1098</td>
</tr>
<tr>
<td>Mar</td>
<td>0.01992</td>
<td>0.006985</td>
<td>0.0044</td>
</tr>
<tr>
<td>Apr</td>
<td>0.01734</td>
<td>0.006922</td>
<td>0.0122</td>
</tr>
<tr>
<td>May</td>
<td>0.01172</td>
<td>0.006866</td>
<td>0.0879</td>
</tr>
<tr>
<td>Jun</td>
<td>0.005451</td>
<td>0.006817</td>
<td>0.4239</td>
</tr>
<tr>
<td>Jul</td>
<td>0.001153</td>
<td>0.006775</td>
<td>0.8649</td>
</tr>
<tr>
<td>Aug</td>
<td>-0.00537</td>
<td>0.006740</td>
<td>0.4257</td>
</tr>
<tr>
<td>Sep</td>
<td>-0.01145</td>
<td>0.006714</td>
<td>0.0882</td>
</tr>
<tr>
<td>Oct</td>
<td>-0.01272</td>
<td>0.006694</td>
<td>0.0575</td>
</tr>
<tr>
<td>Nov</td>
<td>-0.00885</td>
<td>0.006683</td>
<td>0.1855</td>
</tr>
<tr>
<td>Dec</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

### Discussion

#### Summary of Findings.
In this study, a pragmatic, scalable, and CCM-based population health strategy for diabetes care was evaluated in 18 primary care clinics using a before-after study design with an interrupted time series analysis. A key component of the intervention, the point-of-care CDS system, was well accepted, with usage rates rising to approximately 70% of primary care encounters by study participants by the end of the intervention period. Potential reasons for non-universal system use may have included chronic disease care being of secondary concern during clinic visits for acute medical conditions, as well as individual providers making limited use of the tool. Given a usage rate of 88% in one clinic by the end of the study period, and a rate exceeding 80% in five clinics, a system usage target of > 80%, or perhaps even > 90%, appears to be an attainable goal for this type of CDS tool in primary care.

After adjusting for seasonality, the rate of change in A1C levels during the intervention period was -0.02%/year, as compared to 0.01%/year during the historical control period (p = 0.01). However, we do not deem this difference to be clinically significant. Overall, the A1C levels remained remarkably stable during both the pre-intervention period (increasing from 7.17% in January 2007 to 7.20% in December 2008) as well as the intervention period (unchanged, at 7.24% in both January 2009 and December 2010). The overall change in A1C levels over the course of the entire study period, encompassing both the pre-intervention and intervention periods, was 0.07% (7.17% to 7.24%). This change was only 17% of the 0.42% increase expected based on the UKPDS long-term outcomes data and the baseline A1C levels. Thus, the study results suggest that both labor-intensive, disease registry-based population health approaches as well as an EHR-enabled, CCM-based population health program are capable of slowing A1C progression and presumably the associated complications. As noted in the descriptions of these two population health approaches, however, the EHR-facilitated population health program required much less time and resources to operate compared to the stand-alone disease registry-based approach.

In summary, this study contributes insights on the long-term clinical outcomes of an EHR-enabled, team-based, and scalable CCM approach to population health management. Additionally, this study provides validation for the enterprise-level use of standards-based CDS Web services to enable population health management, including for point-of-care CDS and feedback reporting.

#### Limitations and Strengths of Approach.
A limitation of our study is the lack of a concurrent control group, with its associated potential for confounding and the difficulty of definitively attributing the patient outcomes to the study intervention. While clinical leaders were not aware of any other major interventions or changes in clinical practice related to diabetes care during the study period, other confounding factors may have impacted the A1C trends that we observed. To contextualize our findings in the absence of a control group, we estimated A1C progression in the study population using the A1C equation embedded in the UKPDS Outcomes Model. We recognize that this equation is based on the UKPDS long-term outcomes population and has not been validated in a US cohort.
However, we are not aware of published data on long-term A1C progression in patients treated per usual care in the US that we could have used to estimate A1C progression. Another limitation of our approach is the inability to distinguish between the impacts of different components of our multi-faceted intervention. Our inclusion criteria also required patients to be seen at least once a year at the study sites during the study period; therefore, our findings may not be generalizable to patients with diabetes who do not consistently receive primary care. As another limitation, our estimate of time savings from the intervention is based on expert opinion rather than explicit measurement. Also, due to our focus on improving clinical outcomes, we did not evaluate for process measures (e.g., guideline compliance rates), which could have provided more insights into the mechanisms underlying the observed clinical outcomes. Finally, we used A1C results collected as a part of the routine care process, rather than A1C results collected specifically for the study, and at set time points. However, to account for this final limitation, we used interpolation to ensure that all patients had the same number of observations to contribute to the analysis.

A strength is the use of a large sample size, which increases the precision of our findings. In addition, our evaluation encompasses four years of data, which allows for a long-term assessment of the impact of the population health strategies under study. Third, the use of a historical control period provides some level of protection against confounding. Fourth, our system and analyses are population-based and not limited to a subset of poorly controlled patients who might experience substantial A1C improvement with disease management. Finally, the intervention evaluated was pragmatic and scalable, with clear benefits compared to a labor-intensive approach to population health management.

Implications and Future Directions. Overall glycemic control in the study population was quite good both before and after the intervention, with average A1C levels very close to 7%, which is the target goal for many individuals with diabetes. Moreover, A1C levels increased by only 0.07% during the entire evaluation period and remained unchanged during the intervention period, compared to an expected increase of 0.42%. Thus, an important implication of this study is that a team-based, EHR-enabled population health program based on the CCM is capable of effectively managing a population of patients with diabetes. Key components of this population health strategy include the empowerment of care team members, e.g., to order needed laboratory tests by protocol; standard care protocols that harness the efforts of the entire care team, such as the removal of shoes by patients with diabetes during the rooming process to facilitate recommended foot exams; and providing feedback to clinicians to reaffirm the importance of their population health management efforts. While our study design was not capable of teasing out the impact of these various facets of the intervention, our empiric observation is that each of these various facets of the intervention played an important role in the overall effectiveness of the population health program.

Another important implication of our study is that an automated approach to population health management can be just as effective as a highly labor-intensive approach. Indeed, by leveraging the data captured in EHR systems as a part of routine clinical care, an EHR-enabled population health program may be pursued in a highly automated fashion with minimal need for duplicate, manual data entry.

Looking into the future, we anticipate that the increasing adoption of EHR systems, combined with increasing financial pressures to more efficiently and effectively manage chronic conditions, will result in continually greater use of EHR-facilitated, team-oriented, and CCM-based approaches to improving population health. This study contributes to this increasingly important area of biomedical informatics research – the leveraging of information technology to more efficiently manage chronic conditions and improve population health. In the future, we anticipate that many more studies will be conducted by ourselves and others to evaluate how best to improve population health using these technology-enabled and team-based approaches.

Conclusion

An EHR-enabled, team-based, and scalable population health strategy based on the CCM was associated with limited A1C progression among patients receiving outpatient diabetes care. These benefits were comparable to the limited A1C progression associated with a labor-intensive population health management approach requiring extensive manual data entry.

Acknowledgements

The intervention implementation was supported in part by NIH grants R41 LM009051 and R42 LM009051. The authors would like to thank the individuals who contributed to the study, including David E. Shields.
KK is or has been a consultant on CDS for the U.S. Office of the National Coordinator for Health IT, ARUP Laboratories, McKesson InterQual, ESAC, Inc., JBS International, Inc., Inflexxion, Inc., Intelligent Automation, Inc., Partners HealthCare, Mayo Clinic, and the RAND Corporation. KK receives royalties for a Duke University-owned CDS technology for infectious disease management known as CustomID that he helped develop. KK was formerly a consultant for Religent, Inc. and a co-owner and consultant for Clinica Software, Inc., both of which provide commercial CDS services, including through use of a CDS technology known as SEBASTIAN that KK developed, and which was used by the point-of-care CDS system described in this manuscript. KK no longer has a financial relationship with either Religent or Clinica Software. KK also has received grant support from Hitachi. KK has no competing interest with any specific product or intervention evaluated in this manuscript. HB receives grant funding from Takeda, Johnson & Johnson, Improved Patient Outcome, MeadWestVaco and Sanofi. HB also has received consulting funds from Sanofi, Regeneron, CVS, Walgreens, and Blue Cross/Blue Shield of Arkansas. HB has no competing interest with any specific product or intervention evaluated in this manuscript. The other authors have no potential competing interests to declare.

References
i3b3: Infobuttons for i2b2 as a Mechanism for Investigating the Information Needs of Clinical Researchers

Timothy Kennell Jr., BS1,2, Donald M. Dempsey, MS3 and James J. Cimino, MD2

1NIH Medical Scientist Training Program, University of Alabama at Birmingham School of Medicine; 2Informatics Institute and Department of Medicine, University of Alabama at Birmingham; 3Center for Clinical and Translational Science, University of Alabama at Birmingham

Birmingham, Alabama

Abstract
The information needs of clinicians, as they interact with the EHR, are well-studied. Clinical researchers also interact with the EHR and, while they might be expected to have some similar needs, the unique needs that arise due to nature of their work remain largely unstudied. For clinicians, infobuttons (context-aware hyperlinks) provide a mechanism of studying these information needs. Here we describe the integration of infobuttons into i2b2, a popular data warehouse commonly used by clinical researchers, using a plugin. A preliminary survey of i2b2 developers indicates a general interest in infobuttons for i2b2 and indicates good likelihood for their deployment, where they may be used as a tool for further studying these needs in greater detail.

Introduction
Clinical information systems such electronic health records (EHRs), now widely available, address many of the information needs of their users,1,2 yet they simultaneously evoke many new information needs.3,4 The greater availability of EHRs has led to their use by a new constituency: clinical researchers.5,6 Similar to their clinical counterpart, these individuals often provide health care for their subjects as patients and therefore might be expected to have similar information needs. However, this new group of users would also experience different information needs based on their unique use cases, including managing the research workflow of their subject,7 responding to clinical problems that arise because of the research procedures, analyzing the data they collect across their cohort of patients to generate new knowledge,6,8 and using that knowledge to generate new research questions.9,10 It is our hypothesis that these tasks entail information needs that differ from those of routine patient care. However, while the information needs of clinicians have been widely investigated,11–14 an extensive search of PubMed and Google Scholar for studies on those of clinical researches fails to reveal research in this area.

One method for addressing information needs of EHR users is through the context-aware hyperlinks, known as “infobuttons”, that provide patient-specific, just-in-time knowledge at the point of care.15,16 When integrated into EHRs, infobuttons serve not only as a mechanism for decision support, but as a method for studying clinician information needs.17,18 As clinical researchers often access EHR data through clinical data repositories, we believe that integrating infobuttons into these repositories could serve similar dual purposes.

We have therefore developed a method for integrating infobuttons into a popular data warehouse (i2b2). We refer to this version of infobuttons as i3b3. The purpose of this paper is to describe the development of i3b3, explore the interest of the i2b2 community in having such a capability, provide instruction on how i2b2 installations can avail themselves of our software, and discuss the opportunities that i3b3 provides for exploring the varied information needs of clinical researchers.

Background
OpenInfobutton
OpenInfobutton is free-for-use, HL7-compliant infobutton manager provided by the University of Utah (http://www.openinfobutton.org/) that acts as an information broker between infobutton implementations and knowledge resources.14 The EHR infobutton passes relevant, contextual information about the patient, user and user’s status from the originating system using the HL7 standard to OpenInfobutton, which uses that information to select relevant information resources (for example, a drug dosing reference source). It then constructs links to those
resources, customized with the contextual information (for example the drug and patient age and gender). The links are presented back to the user, who can then select those that appear to address his/her information need.

**LITE**

The selection of resources by OpenInfobutton, and the customization of links to those resources, is made possible by a knowledge base used by OpenInfobutton that contains institution-specific information provided by an analyst or librarian responsible for the infobuttons in the institution’s EHR. This customization is provide be a tool called the Librarian Infobutton Manager (LITE), which is also freely available from the University of Utah (http://lite.bmi.utah.edu).

**i2b2**

The i2b2 (Informatics for Integrating Biology and the Bedside) system (https://www.i2b2.org/) is a popular patient data warehouse consisting of a repository for storing patient data and a user interface for rapidly querying the data for research cohorts that meet pre-selected criteria. This system was built by separating individual components into “cells” (interoperable modules) of a larger “hive” that communicate with each other via web protocols. This architecture allows the system to be highly extensible through the development of additional cells and plugins in order add desired functionality. In order to interact with the data repository in the cells, i2b2 provides two types of user interfaces (UIs): a workbench and a webclient. The workbench is a Java-based thin client that requires local installation and configuration on each computer that needs to interact with i2b2. The more common webclient is a browser-based UI built with standard web technologies that interacts with i2b2 via a web proxy server.

**Methods**

**i2b2 Plugin Development**

i2b2 can installed on many different server architectures and it’s functionality can easily be extended through cell or plugin development. Cell development allows the addition of new server-side functionality to i2b2 through communication with other cells in the hive, and can allow additional data to be retrieved directly from the data repository. While this provides extensive flexibility over data manipulation at the database level, cell installation in currently implemented i2b2 instances requires substantial modification to server configuration files. Additionally, user interaction with a custom i2b2 cell typically requires the development of an externally-facing plugin on the webclient as well.

Because we desired to make installation of our i3b3 infobuttons as simple as possible, we chose to develop a plugin. These can be installed into i2b2 by simply adding the plugin’s directory structure to a standard location in the webclient’s web server and by changing a single configuration file.

In addition to i2b2’s standard plugin structure, we connected two other systems to it in order to establish communication between i2b2 and external knowledge resources. Briefly, i2b2’s data are sent via a HL7-compliant HTTP GET web request to OpenInfobutton. OpenInfobutton then generates a set of links to external resources based on prior settings configured in LITE. The links are then returned to the user as a list of uniform resource locations (URLs) displayed in a browser window for user interaction.

**Information Needs Survey**

In order to make a preliminary determination of the general interest in the integration of infobuttons into i2b2, we constructed a Google survey (http://goo.gl/forms/Xulqq5s3QW) (see Figure 1). As the survey was a preliminary information needs assessment, we kept the survey short and briefly reviewed the two included questions before finalizing it. The first question was simply, “Would you be interested in infobuttons for i2b2?” (yes/no). The second question was only revealed to participants if they responded “Yes” to the first question. In order to provide participants with a framework for possible locations that infobuttons could be integrated into i2b2, this question began with a screenshot of the i2b2 plugin that we developed. The question following the image, “Where would you like to integrate infobuttons into i2b2 (choose as many as apply):”?, then provided both a list of checkboxes of logical locations to integrate infobuttons.

After completing and reviewing the survey, we emailed generated link to the survey to all recipients on the i2b2 mailing list and posted the link on the “i2b2 Install Help” Google group.
We collected responses for the following 10 days followed by creating a summary of the data collected.

### Question 1

**Interest in Infobuttons for i2b2**

Infobuttons are context-aware links that have been integrated into EHRs for retrieval of information from external resources. Their ability to perform context-aware searches allows them to perform faster searches for more relevant information than a simple web search for the same information. While infobuttons have been integrated into EHRs in the past, they may provide a similar benefit for informational retrieval regarding patient sets if integrated into research data management systems such as i2b2.

* Required

Would you be interested in infobuttons for i2b2? *

- Yes
- No

### Question 2

**Where would you like to integrate infobuttons into i2b2 (choose as many as you want):** *

- Diseases
- Medications
- Laboratory Results
- Demographics
- Procedures
- Imaging
- Radiology
- Pathology
- Clinical Notes
- Navigation Terms
- Previous Queries
- Other:

Figure 1. Information needs survey for assessing infobutton interest in the i2b2 community and desired locations for integrating infobuttons located at [http://goo.gl/forms/XuIqq5s3QW](http://goo.gl/forms/XuIqq5s3QW). The first question (Question 1) simply asks the respondent if they are interested in having infobuttons in i2b2 (yes/no). If they answer “yes”, the second question allows the respondent to detail where they would like infobuttons to be integrated. At the beginning, question 2 shows a simple screenshot of the plugin’s UI for reference and then provides several locations for integration including several concept classes present in i2b2 and a fill-in-the-blank response for alternative suggestions.

### Results

#### Plugin Development and Use

Integration of infobuttons into the i2b2 system combines the plugin system, OpenInfobutton, and LITE. The plugin consists of two main components: a UI, built with HTML5 and CSS3, and a controller, built with JavaScript and Prototype (a JavaScript framework), each constructed according to the i2b2 plugin standard [https://community.i2b2.org/wiki/display/webclient/Web+Client+Plug-in+Developers+Guide](https://community.i2b2.org/wiki/display/webclient/Web+Client+Plug-in+Developers+Guide). Figure 2 provides a schematic view of the components and processes involved.
After selecting the plugin from i2b2’s list (see Figure 3), a user can select a concept class, such as *diseases* or *medications*, to investigate in a patient set and then drag a previously created patient set into a text box in the plugin window. This action will cause the controller to generate an AJAX (asynchronous JavaScript and XML) request directed toward the i2b2 data repository, which will respond with an XML message containing the requested information. After receiving the response, the controller parses through it to prepare it for display to the user. Once finished, the controller updates the UI with a list of the patients in the set and the respective information requested (i.e. a list of diseases or medications for each patient; see Figure 4).

![Image](image.png)

**Figure 2. i3b3 plugin architecture.** Initially, a i2b2 user creates a patient set in i2b2 and loads the i3b3 plugin. The user can then drag the previously created patient set into the i3b3 frame (1) causing a list of patients and associated concepts (medications and diagnoses) to be displayed. Clicking on an infobutton next to any of the concepts opens a secondary browser window (resource window) and sends context-specific information to OpenInfobutton (2). OpenInfobutton queries to its resource knowledge base (created using LITE - the Librarian Infobutton Tailogin Environment) and generates a list of customized hyperlinks (3) that link to selected resources (4). The user can select one or more of the available resource links in the resources window to receive information to their initial question (5).

Next to each of the concepts listed below each patient, the controller places an infobutton icon that contains a URL with HTTP GET parameters to be passed to OpenInfobutton (see Figure 4). The parameters in the URL, which include the displayed concept, the patient’s age, and the patient’s gender as well as the institution (“UAB”) and the task (“i2b2”), are context-aware and translated from i2b2’s terminology into HL7-compliant terms. A user’s click on one of the infobutton icons sends a request to OpenInfobutton containing the previously mentioned information. The infobutton manager will then utilize a pre-configured, XML settings file (created with LITE) to select appropriate resources and construct links to those resources that include the appropriate parameter values obtained from the plugin. The links to each resource are combined into a final output in a secondary window. Each resource returns a result set on the original concept automatically refined by the concept’s original context. The result is context-specific links to knowledge resources based on the original concept that was selected (see Figure 5).
Figure 3. Initial state of the i3b3 plugin after loading. A user can select the plugin from the list of i2b2 plugins in the “analysis tools” section [(A); shown enlarged in (B)]. The plugin is then loaded into the right panel (C) with brief instructions displayed at the top and more detailed instructions in a “Help” tab located at the top of the plugin panel.

Figure 4. Example of user interaction with i3b3 and the response provided. (1) A user selects a class of concepts that he/she desires to know more about from an available list (example depicts “Medications” chosen). (2) The user then selects a patient set from a previous query (shown enlarged in “A”) and (3) drags and drops it into the available space in the plugin panel. (4) i3b3 will then return a list of the patients in the set and the associated concepts initially chosen in step (1). Each concept (in this case, medications, shown enlarged in “B”) has an associated infobutton that links to OpenInfobutton.
The finalized infobutton plugin has been distributed via our institution’s GitHub page found at https://github.com/uab-informatics-institute/i3b3. This release method, combined with the design of the system as an i2b2 plugin, allows for rapid installation in any i2b2 instance requiring only minimal instructions that are provided on the GitHub page. Additionally, GitHub provides a simple method for community contributions to future development of the plugin. Following construction and distribution, the plugin was easily installed by a system administrator in our institution’s i2b2 development environment with only the GitHub URL and confirmation of the i2b2 concept pathway structure provided. This highlights the fact that the i3b3 plugin remains independent of the actual ontology used in an i2b2 installation and needs only a minor setting modification in order to provide the location specific classes of concepts desired to be present for user selection.

Survey Responses
The distribution method of the survey allowed it to reach a wide audience from new to experienced researchers and clinicians along with several system analysts responsible for installing and maintaining i2b2 at their institution. The survey reached approximately 285 individuals through the “i2b2 Install Help” Google group and an unknown number through the mailing list. In total, the survey received 19 responses with all answering “Yes” to the question about interest in infobuttons for i2b2 (see Figure 6). In response to the question regarding desired infobuttons domains, participants requested infobutton integration for Diseases, Medications, Laboratory Results, and Procedures most frequently while others received a much lower frequency of requests. This trend may indicate the greatest, unmet information needs that arise in clinical researchers during interaction with i2b2.

Most respondents did not take advantage of the “Other” category, suggesting that the majority of potential locations for infobutton integration were already present in the survey options. Of those participants that did choose the
“Other” category, one simply suggested that all of the categories listed as answer choices would be useful and added that an infobutton for genotypic data would also be useful. Taken together, these results suggest the presence of unmet information needs in the i2b2 community, potentially extending to clinical investigators. Additionally, the survey responses indicate that these information needs may be further studied and potentially met with the use of infobuttons in applications that clinical investigators frequently interact with during the course of their studies.

Figure 6. Results of the i2b2 infobutton survey. There were 19 total responses to the survey. All respondents answered “Yes” to the first question asking about interest in infobutton for i2b2. The figure above shows the responses to the second question asking about location integration. General concepts such as diseases, medications, laboratory results, and procedures received the highest recommendations with most other options chosen much less frequently.

Discussion
This paper describes an easy-to-adopt solution to the integration of infobuttons into any i2b2 installation. Coupled with the freely available OpenInfobutton and LITE, any i2b2 instance in the world can now provide its users with context-aware knowledge resource links and, based on our survey, it seems there is significant interest in doing so.

What links to provide remains an open question. The literature has few studies investigating the information needs of clinical scientists, and managers of i2b2 systems should not be quick to assume that the needs of clinical researchers will be the same as those of clinicians. More studies of clinical researchers’ needs will be required as studies of clinicians’ needs consistently show that those needs are task-dependent. While clinical researchers may have similar needs because they may have similar tasks (that is, caring for patients in their studies), they have additional tasks related to research, including adherence to study protocol, analysis and publication of data, and new hypothesis generation, all of which may be triggered by the same EHR data that trigger clinician needs. Those studies also consistently show that information needs cannot always be predicted without empiric observation. We conclude, therefore, that the information needs of clinical researchers should be a target of future study as part of the deployment of infobuttons. As with studies of clinician needs, infobuttons themselves (in this case, i3b3) can serve as a valuable tool in such studies.

The plugin we constructed was only developed and tested within the purview of a single hospital. However, we do not believe that our hospital environment and clinical research environment differ from other hospitals in ways that would impact i2b2 implementation, and we believe that our experience is applicable to many other medical institutions. Although our integration of infobuttons into i2b2 through a plugin was limited in the number of
contexts in which information needs might be addressed, the results of our survey show that the initial scope chosen for the plugin covers some of the most important contexts, according to the i2b2 developer community. Furthermore, the design of the plugin easily allows for the scope to be extended in order to cover or exclude those information needs deemed important or irrelevant for future study respectively. While the plugin was developed for installment into a single data management system, i2b2, and might not be able to be easily installed in a different data warehouse, i2b2 is a popular option for installation at many hospital systems and i2b2 is not dissimilar to other data management systems that are available. Additionally, while the plugin does not incorporate group metrics for its context-aware searches, we believe that i3b3 provides an initial step in studying the clinical scientists information needs with group metrics being an additional feature that could be added as needed. Therefore, the results of this study, including the infobutton plugin, should be applicable to research institutions without i2b2 instances.

**Conclusion**

We have developed software that enables interested i2b2 developers (of which there are at least 19) to provide infobuttons to their users. OpenInfobutton and LITE, also freely available, provide a comprehensive infobutton management strategy. What to do with infobuttons in i2b2 contexts is currently unknown, but the installation of i3b3 provides a mechanism for studying this question, much as infobuttons have been used to study information needs in clinical settings.

**Acknowledgments**

Mr. Kennell is supported by an NIH Medical Student Training Program grant to the University of Alabama at Birmingham under grant ST32GM008361-23 and a TL1 grant from the University of Alabama School of Medicine (UASOM) Center for Clinical and Translational Sciences (CCTS) under grant 1TL1TR001418-01 from the National Center for the Advancement of Translational Science (NCATS). Dr. Cimino is also supported by the CCTS NCATS grant and by research funds from the UASOM Informatics Institute. The authors thank R. Dale Johnson, BS and Matt C. Wyatt, MS for their advice on plugin development and Guilherme Del Fiol.

**References**

Feasibility of Extracting Key Elements from ClinicalTrials.gov to Support Clinicians’ Patient Care Decisions

Heejun Kim, MS¹, Jiantao Bian, MS², Javed Mostafa, PhD¹, Siddhartha Jonnalagadda, PhD³, Guilherme Del Fiol, MD, PhD²

¹School of Information and Library Science, University of North Carolina, Chapel Hill, NC, USA; ²Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA; ³Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Abstract

Motivation: Clinicians need up-to-date evidence from high quality clinical trials to support clinical decisions. However, applying evidence from the primary literature requires significant effort.

Objective: To examine the feasibility of automatically extracting key clinical trial information from ClinicalTrials.gov.

Methods: We assessed the coverage of ClinicalTrials.gov for high quality clinical studies that are indexed in PubMed. Using 140 random ClinicalTrials.gov records, we developed and tested rules for the automatic extraction of key information.

Results: The rate of high quality clinical trial registration in ClinicalTrials.gov increased from 0.2% in 2005 to 17% in 2015. Trials reporting results increased from 3% in 2005 to 19% in 2015. The accuracy of the automatic extraction algorithm for 10 trial attributes was 90% on average. Future research is needed to improve the algorithm accuracy and to design information displays to optimally present trial information to clinicians.

Introduction

Evidence Based Medicine (EBM) is the application of the best and up-to-date scientific evidence available from rigorous clinical studies to guide the care of individual patients. To practice EBM, clinicians need to retrieve, appraise, and integrate the latest research findings into decisions for a particular patient. However, clinicians often cannot process the large amount of recent research results within the typical busy patient care environment. As a result, over 60% of the clinical questions that clinicians raise go unanswered due to the amount of time needed to spend on the search task. Systematic reviews are a common approach to synthesize the evidence on a specific clinical topic, but the process is slow and expensive. As a result, many clinical topics have no systematic reviews and a large percentage of published reviews are outdated. Hence, solutions are needed to help clinicians use the results of original clinical studies in the care of their patients.

To improve the efficiency of information seeking, substantial progress has been made to improve the evidence retrieval process. Several approaches are available to help clinicians identify high quality and high impact clinical studies. However, less attention has been dedicated to help clinicians quickly judge the relevance of a clinical study to a particular patient and understand the gist of the study findings. To improve the search process, clinicians have been suggested to formulate their questions according to four components: patient population, intervention, comparison, and the outcome of interest (PICO). While several information retrieval studies have used the PICO framework, most prior work has focused on the search process. Preliminary studies have demonstrated the feasibility of automatically extracting core clinical trial characteristics and results from narrative text such as journal article abstracts. There have also been several studies that extracted PICO elements from narrative text. More recently, the increased use of clinical trial registries that record semi-structured trial information provides a unique opportunity for improving the information extraction process.

The overall goal of the present study is to assess the feasibility of automatically extracting PICO elements of clinical trials from ClinicalTrials.gov. Specifically, we aimed to (1) assess the coverage (i.e., the rate of high quality clinical trials indexed in PubMed that have been registered in ClinicalTrials.gov) of ClinicalTrials.gov for high quality trials to examine its comprehensiveness as a data source; and (2) develop and assess the accuracy of an algorithm to extract PICO elements from ClinicalTrials.gov that overcomes unexpected patterns of registration. Both coverage and
extraction algorithm accuracy are important for the feasibility of extracting PICO information from ClinicalTrials.gov. The resulting algorithm can be used as a component of tools that support clinicians’ decision making.

Background

ClinicalTrials.gov is an online registry of clinical research trials established by the US National Library of Medicine (NLM) in 1999 aiming for increased transparency in clinical research. ClinicalTrials.gov archives over 209 thousand trials from 192 countries as of February 2016. Since 2007, trials covered by the Food and Drug Administration Amendments Act (FDAAA) are required to be updated within one year of completion with basic summary results data. A ClinicalTrials.gov record contains detailed information about a clinical trial in semi-structured XML format. Trial records include information such as conditions, eligibility, sample size, study methods, study arms, and study outcomes. In the present study, we leverage the semi-structured information from ClinicalTrials.gov to automatically extract PICO elements.

Links between Medline and ClinicalTrials.gov. When registering a trial on ClinicalTrials.gov, a unique identifier (“NCT”) is assigned automatically by the registry. The NCT identifier is also recorded in the Medline citation of articles that report the results of a registered trial, establishing a link between Medline citations and ClinicalTrials.gov records. The International Committee of Medical Journal Editors (ICMJE) has required that all clinical trials be publicly registered before manuscripts are submitted and accepted for publication. ClinicalTrials.gov provides two types of XML files: general information about the trial methods and the trial results. These files can be retrieved through a RESTful (Representational State Transfer) Web service, which was used in our study method.

Huser and Cimino examined links between ClinicalTrials.gov records and PubMed citations for trials completed between 2006 and 2013. They found that 28% of the trial records had at least one linked article. The study also found that 27% of the trials completed by 2009 had basic results. In our study, we looked at the reverse, i.e. links from clinical trial citations in PubMed to ClinicalTrials.gov. For clinical decision support purposes, it is more reasonable to start the information seeking process by searching for relevant high quality trials in PubMed than by searching for trials records in ClinicalTrials.gov.

Previous studies extracting PICO elements. In a systematic review, Jonnalagadda et al. found four studies that automatically extracted PICO elements from biomedical publications. Huang et al. used a naïve Bayes classifier, Boudin et al. used a combination of multiple classifiers (random forest, naïve Bayes, support vector machines, and multi-layer perceptron), and Demner-Fushman and Lin utilized a rule-based approach to detect PICO elements in medical abstracts. Boudin et al. utilized features such as Medical Subject Headings (MeSH) semantic types, word overlap with article title, and number of punctuation marks, while Huang et al. used terms with highest frequency. The average precision over PICO elements for these studies ranged from 0.63 to 0.84.

Method

The study method consisted of: (1) retrieving citations of recent high quality trials from PubMed and examining coverage of ClinicalTrials.gov for high quality trials; (2) retrieving ClinicalTrials.gov records for the trials retrieved in Step 1; (3) selecting 40 random ClinicalTrials.gov records from the set in Step 2 and manually inspecting the PICO elements from the records of these trials; (4) writing a set of rules to extract PICO elements from trial records; and (5) evaluating the accuracy of the rules with an independent set of 100 trials.

Retrieving citations of recent high quality trials from PubMed. To identify high quality clinical studies in PubMed, we utilized the machine learning classifier developed by Kilicoglu et al. This classifier uses a Naïve Bayes algorithm and utilizes features such as word frequency, semantic features, and citation metadata from PubMed. The precision of the classifier is 83%. We retrieved 178,302 citations of high quality studies from PubMed published in high impact journals between December 2005 and December 2015 by applying the classifier. Next, we used the E-utils API to retrieve full PubMed citations in XML format. A Python script was used to extract the ClinicalTrials.gov identifier (NCT) from the <AccessionNumber> tag where the value of the parent <DataBankName> tag was “ClinicalTrials.gov” (Figure 1).
For instance, the following URL can be used to retrieve the PubMed citation in XML file for the article with PubMed identifier 24725238: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=24725238&report=xml. The coverage of ClinicalTrials.gov for high quality trials was calculated by dividing the number of PubMed citations with a link to ClinicalTrials.gov by the total number of high quality citations. Annual coverage rates were obtained for the 2005 to 2015 period.

Retrieving ClinicalTrials.gov records. We used the ClinicalTrials.gov Web service API to retrieve trial records in XML format. For each trial, we retrieved two types of XML files using a Python script: general information about the trial methods and the trial results. For instance, the methods record for the trial with identifier “01768286” was retrieved using: http://clinicaltrials.gov/ct2/show/NCT01768286/?displayxml=true. The trial results record was retrieved with: https://clinicaltrials.gov/ct2/show/NCT01768286/?resultsxml=true (Figure 2). There is a temporal gap between registration of clinical trials methods and reporting their results, so there are cases in which the registered trials do not include results. Only trials that have been completed with basic summary results were selected for the extraction study. We also calculated the percentage of those ClinicalTrials.gov records that reported trial results.

Selecting ClinicalTrials.gov records. From the trial records retrieved in the previous step, we selected a random set of 40 records (development set) out of 463 trials completed in 2014 to develop a set of rules for PICO extraction. An independent random set of 100 records (evaluation set) was selected from the same 2014 set for evaluating the accuracy of the algorithm.

PICO extraction rules. One of the authors (HK), assisted by two co-authors (GDF, JM), manually reviewed the 40 trial records in the development set. Based on this review, rules were developed to automatically extract PICO elements from the trial records. We referred to the ClinicalTrials.gov XML schema (https://clinicaltrials.gov/ct2/html/images/info/public.xsd) to guide rule development and to understand how ClinicalTrials.gov users are instructed on registering trials.

Rules were designed to extract 10 trial attributes associated with PICO elements (Table 1). The rules are basically a combination of XPATH expressions. For example, from the <condition> tag (Case 1, Table 2), <minimum_age> tag (Case 2, Table 2), and <enrollment> tag (Case 3, Table 2), we extracted data elements related to the study population, i.e. the main condition of interests in study participants, minimum participant age, and number of total participants in the study inclusion criteria.

Figure 1. Fragment of a Medline citation in XML format retrieved using the NCBI E-utilities. The squares highlight the relevant nodes for the extraction of ClinicalTrials.gov identifiers.

For instance, the following URL can be used to retrieve the PubMed citation in XML file for the article with PubMed identifier 24725238: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=24725238&report=xml. The coverage of ClinicalTrials.gov for high quality trials was calculated by dividing the number of PubMed citations with a link to ClinicalTrials.gov by the total number of high quality citations. Annual coverage rates were obtained for the 2005 to 2015 period.

Retrieving ClinicalTrials.gov records. We used the ClinicalTrials.gov Web service API to retrieve trial records in XML format. For each trial, we retrieved two types of XML files using a Python script: general information about the trial methods and the trial results. For instance, the methods record for the trial with identifier “01768286” was retrieved using: http://clinicaltrials.gov/ct2/show/NCT01768286/?displayxml=true. The trial results record was retrieved with: https://clinicaltrials.gov/ct2/show/NCT01768286/?resultsxml=true (Figure 2). There is a temporal gap between registration of clinical trials methods and reporting their results, so there are cases in which the registered trials do not include results. Only trials that have been completed with basic summary results were selected for the extraction study. We also calculated the percentage of those ClinicalTrials.gov records that reported trial results.

Selecting ClinicalTrials.gov records. From the trial records retrieved in the previous step, we selected a random set of 40 records (development set) out of 463 trials completed in 2014 to develop a set of rules for PICO extraction. An independent random set of 100 records (evaluation set) was selected from the same 2014 set for evaluating the accuracy of the algorithm.

PICO extraction rules. One of the authors (HK), assisted by two co-authors (GDF, JM), manually reviewed the 40 trial records in the development set. Based on this review, rules were developed to automatically extract PICO elements from the trial records. We referred to the ClinicalTrials.gov XML schema (https://clinicaltrials.gov/ct2/html/images/info/public.xsd) to guide rule development and to understand how ClinicalTrials.gov users are instructed on registering trials.

Rules were designed to extract 10 trial attributes associated with PICO elements (Table 1). The rules are basically a combination of XPATH expressions. For example, from the <condition> tag (Case 1, Table 2), <minimum_age> tag (Case 2, Table 2), and <enrollment> tag (Case 3, Table 2), we extracted data elements related to the study population, i.e. the main condition of interests in study participants, minimum participant age, and number of total participants in the study inclusion criteria.
Table 1. Trial attributes extracted by the PICO extraction algorithm for each PICO element.

<table>
<thead>
<tr>
<th>PICO elements</th>
<th>Trial attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Main condition of interest in study participants, minimum and maximum participant age, and total enrollment</td>
</tr>
<tr>
<td>Intervention and comparison</td>
<td>Number of study arms and study arm interventions</td>
</tr>
<tr>
<td>Outcome</td>
<td>Number of participants who started the trial, number of participants who completed the trial, primary outcome, and results for the primary outcome.</td>
</tr>
</tbody>
</table>

The rules retrieved scattered PICO elements over two XMLs from several tags and merged them based on the study arm intervention. For instance, the number of participants who completed the trial is only available from the `<participants>` tag (Case 5, Table 2) under the `<participant_flow>` tag. However, the study arm design is often changed and needs to be verified from the `<title>` tag (Case 4, Table 2) under the `<outcome_list>` tag if it is the final study arm intervention. If there is discrepant information (e.g., pre-arm design before trials and real-arm intervention), the latest information is selected.

We retrieved outcome elements from the `<clinical_results>` tag. Description for the primary outcome measure was extracted from the `<title>` tag under the `<outcome>` tag. The actual outcome measurements were retrieved from the `value` attribute of the `<measurement>` tag as in Case 6 (Table 2). The algorithm produces an output in XML format according to the XML schema available at (https://sites.google.com/site/automaticpicoextraction/xml-schema). The complete rules set is available at (https://sites.google.com/site/automaticpicoextraction).
**Table 2.** Examples of PICO elements successfully extracted from clinical trial records in ClinicalTrials.gov and corresponding XPATH.

<table>
<thead>
<tr>
<th>Case</th>
<th>Identifier (NCT)</th>
<th>Element type</th>
<th>Sample XML node</th>
<th>Pseudo code with XPATH</th>
<th>Extracted result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>00006237</td>
<td>Main condition of interests in study participants</td>
<td><code>&lt;condition&gt;Melanoma (Skin)&lt;/condition&gt;</code></td>
<td><code>//clinical_study/condition</code></td>
<td>Melanoma (Skin)</td>
</tr>
<tr>
<td>2</td>
<td>00006237</td>
<td>Minimum participant age</td>
<td><code>&lt;minimum_age&gt;18 Years&lt;/minimum_age&gt;</code></td>
<td><code>//clinical_study/eligibility/minimum_age</code></td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>00006237</td>
<td>Total enrollment</td>
<td><code>&lt;enrollment type=&quot;Actual&quot;&gt;432&lt;/enrollment&gt;</code></td>
<td><code>//clinical_study/enrollment[@type=&quot;Actual&quot;]</code></td>
<td>432</td>
</tr>
<tr>
<td>4</td>
<td>00081731</td>
<td>Study arm intervention</td>
<td><code>&lt;group group_id=&quot;O1&quot;&gt;&lt;title&gt;Optimal Medical Therapy&lt;/title&gt;&lt;/group&gt;</code></td>
<td><code>//clinical_study/clinical_results/outcome_list/outcome[type=&quot;Primary&quot;]/group_list/group/title</code></td>
<td>Optimal Medical Therapy</td>
</tr>
<tr>
<td>5</td>
<td>00081731</td>
<td>Number of participants who completed the trial</td>
<td><code>&lt;participant_flow&gt;…&lt;group group_id=&quot;P1&quot;&gt;&lt;title&gt;Optimal Medical Therapy&lt;/title&gt;&lt;/group&gt;…&lt;participants group_id=&quot;P1&quot; count=&quot;472&quot;&gt;…&lt;/participant_flow&gt;</code></td>
<td><code>//clinical_study/clinical_results/participant_flow/period_list/period/milestone_list/milestone[@title=&quot;COMPLETED&quot;]/participants_list/participants[@count]</code></td>
<td>472</td>
</tr>
<tr>
<td>6</td>
<td>00081731</td>
<td>Results for the primary outcome</td>
<td><code>&lt;measure&gt;…&lt;measurement group_id=&quot;O1&quot; value=&quot;20&quot;/&gt;&lt;/measurement_list&gt;</code></td>
<td><code>//clinical_study/clinical_results/outcome_list/outcome[type=&quot;Primary&quot;]/group_list/group/title</code></td>
<td>20 (Cardiovascular or Renal Death)</td>
</tr>
</tbody>
</table>

**Evaluation.** Two of the co-authors (JB, GDF) rated the algorithm output for the 10 attributes described above. To determine the accuracy of each extracted attribute, the raters compared the algorithm output with the original trial record at the ClinicalTrials.gov Web site. An attribute was considered to be accurate when the attribute value in the algorithm output was identical to the value that could be manually searched in ClinicalTrials.gov. After an initial round of calibration with 20 trials, the Cohen’s Kappa inter rater agreement was 0.51. Disagreements were resolved by consensus between the two raters. For instance, in one case there was disagreement on successful extraction of study arm interventions. The algorithm correctly identified three out of five arm interventions. One author considered it as accurate because three arms were extracted correctly, but the other author considered it inaccurate due to incompleteness. We agreed that the algorithm is expected to extract all the arms correctly. With the calibrated evaluation criteria, one of the evaluators (JB) continued the evaluation on the remaining 80 trials. For one attribute (number of participants who completed the trial), there were 9 trials that did not have data in ClinicalTrials.gov. They were excluded from the evaluation. In total, 991 trial attributes were evaluated.

**Results**

**ClinicalTrials.gov coverage.** Out of 178,302 high quality clinical studies indexed in PubMed between 2005 and 2015, 15,360 (8.6%) had a record in ClinicalTrials.gov. Among those registered trials, 3,496 (22.7%) trials had basic results. The coverage of ClinicalTrials.gov for high quality clinical studies has increased from 0.2% in 2005 to 17% in 2015 (Figure 3). The percentage of trials in ClinicalTrials.gov that included results has increased from 3% in 2005 to 19% in 2015. The rate of coverage for trial results peaked (28%) in 2012, but decreased afterwards.
Algorithm accuracy. Out of 991 trial attributes extracted from 100 trial records, 893 (90%) could be accurately extracted. Overall, the population-related elements (94%) and intervention- and comparison-related elements (97%) were relatively easy to extract compared to the outcome-related elements (83%). The algorithm extracting the total enrollment hit the highest accuracy (100%) and the algorithm extracting the number of participants who started the trial showed the lowest accuracy (70%). Table 3 provides the accuracy for each trial attribute.

Table 3. Accuracy of the PICO extraction algorithm according to 10 trial attributes for 100 trial records.

<table>
<thead>
<tr>
<th></th>
<th>Main condition of interest in study participants</th>
<th>Minimum participant age</th>
<th>Maximum participant age</th>
<th>Total enrollment</th>
<th>Number of study arms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>77 (77%)</td>
<td>99 (99%)</td>
<td>99 (99%)</td>
<td>100 (100%)</td>
<td>96 (96%)</td>
</tr>
<tr>
<td>Study arm interventions</td>
<td>Number of participants who started the trial</td>
<td>Number of participants who completed the trial</td>
<td>Primary outcome</td>
<td>Results for the primary outcome</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>97 (97%)</td>
<td>70 (70%)</td>
<td>72 (79%, 72/91)</td>
<td>97 (97%)</td>
<td>86 (86%)</td>
</tr>
</tbody>
</table>

Discussion

In this study, we aimed to assess the coverage of ClinicalTrials.gov for high quality trials indexed in PubMed and to examine the feasibility of extracting PICO elements from ClinicalTrials.gov records. To our knowledge, this is the first attempt to examine the feasibility of extracting PICO elements from ClinicalTrials.gov to support clinicians’ information needs. According to our analysis, although the overall coverage of ClinicalTrials.gov is still relatively low, the coverage has steadily increased since 2005 and clinical trials with higher impact should be more likely to be registered due to regulatory and publication requirements. For trials that are registered, most PICO elements could be accurately extracted from the randomly selected 100 ClinicalTrials.gov records using simple XPATH-based rules. Our proposed method could be used to enable alternate information displays to summarize evidence from clinical trials in a way that reduces clinicians’ cognitive effort. For example, in a study by Slager et al., physicians rated an information display structured according to PICO elements more favorably when compared with PubMed narrative.
The algorithm developed in the present study can be used to support the implementation of such structured displays.

Several studies have attempted to extract PICO elements with machine learning algorithms and/or simple rules. Previous approaches were assessed in terms of precision (63-84%) because PICO elements were extracted from narrative texts and therefore the methods were framed as information retrieval problems. However, our approach aimed to extract PICO elements from semi-structured data, with only one extraction target per data element. Thus, we used accuracy (90%) for measurement. As a result, direct comparisons with previous studies are difficult. The level of extraction among studies was also different. Three of those studies classified PICO elements at the sentence-level, while one study extracted PICO elements at the object-level like ours. Overall, our rule-based approach showed fairly good performance most likely because previous approaches relied on the extraction of information from narrative abstracts. However, unlike ours, previous methods do not depend on trial registration in ClinicalTrials.gov and work with any clinical trial published in PubMed. An optimal solution could be to combine both methods, i.e. PICO extraction from ClinicalTrials.gov whenever a trial record is available and extraction from PubMed otherwise.

Extracting the number of participants enrolled in a clinical trial was the easiest task, since the actual sample size was consistently recorded in ClinicalTrials.gov. Overall, elements related to the population and study arm intervention were also easy to extract. Other tasks for extracting main condition of participants, number of participants who started/completed, and results for primary outcomes required careful considerations for unexpected patterns of registration. After accounting for most of the variation in our sample, the algorithm accuracy for those challenging cases was fairly high (70% to 86%).

The main challenge of applying the proposed approach is the low rate of high quality trials that are registered on ClinicalTrials.gov and variation in trial registration format and content. The coverage of ClinicalTrials.gov (11.2%) and the rate of trials with results (21.3%) are fairly low in 2014, however, they have increased gradually likely due to increased regulatory pressure. Moreover, some systematic delays are expected in registering the summary result, as trials covered by the Food and Drug Administration Amendments Act (FDAAA) are required to be updated within one year of completion. If there is a certain degree of delay, PICO elements from ClinicalTrials.gov will not be available in a timely manner for recent trials. An analysis of common algorithm errors is described below.

**Algorithm error analysis.** First, the rules extracted only one condition regardless of the multiple conditions for study participants registered in ClinicalTrials.gov. This error can be fairly easy to correct so that higher accuracy is expected for the next phase of the development. However, there were some cases in which registered information was incomplete. For instance, “Opioid-Induced Constipation (OIC)” was the original condition registered in the ClinicalTrials.gov and extracted correctly by the rules, but our evaluator (JB) found “Non-cancer-related Pain” as an additional condition by reading other tags (Case 7, Table 4).

Second, there were several time series measurements in the results for primary outcome, but the rules could not extract them properly. In Case 8 (Table 4), there were 11 measurements registered in the time series, but the rules extracted only the first. In the Table 4, three measurements were only included for explanation. Corresponding enrollment information on each study arm per each point of measurement were recorded in `<period_list>` tag. The revised rules should process those multiple measurements.

Finally, outcome related attributes such as number of participants who started/completed and results for the primary outcome were the most challenging task while extracting PICO elements from ClinicalTrials.gov. As mentioned earlier, there are two types of XML files in ClinicalTrials.gov: general information about the trial methods and the trial results. There is a temporal gap between inputs for two XMLs. Usually, the former is input before the actual trial, and the latter is input after the actual trial. During this period, the arm design can be changed for several reasons. Thus, there can be different arm information in two XMLs. Moreover, there are several tags containing arm related information such as `<participant_flow>` tag and `<outcome_list>` tag. Unfortunately, users of ClinicalTrials.gov often input data inconsistently so that study arm related information becomes challenging to extract.

As in Case 9 (Table 4), there were six cases in which the values of tags for study arm interventions were changed completely. In the beginning of the study, all participants in the study were to be treated with directional atherectomy. However, primary patency rate was measured with patients treated for Claudication RCC 1-3 in the real study. Since the number of participants who completed the trial is only recorded in the `<participant_flow>` tag and the final intervention design is only recorded in `<outcome_list>` tag, there is no way to match the study arm intervention with the number of participants who completed the trial. ClinicalTrials.gov needs to consider a way to alert its users when
they register a different study arm intervention from the study plan so that they can update misinformation accordingly based on the changed study design. There were also nine cases in which the arm names were slightly changed. For instance, in the trial of NCT00211887, the arm name “Interferon Beta 1a” was changed to an acronym, “IFN-1a.” The next phase of the development should incorporate advanced text extraction algorithms, such as named-entity recognition (NER) and medical semantic networks such as Unified Medical Language System (UMLS) to overcome the current limitation of the rules.

Table 4. Examples of challenging cases for extracting PICO elements from clinical trial records in ClinicalTrials.gov.

<table>
<thead>
<tr>
<th>Case</th>
<th>NCT identifier</th>
<th>Element type</th>
<th>XML node for challenging case</th>
<th>Number of similar cases</th>
<th>Type of error</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>01336205</td>
<td>Main condition of interests in study participants</td>
<td><code>&lt;condition&gt;Opioid-Induced Constipation (OIC)&lt;/condition&gt;</code>&lt;br&gt;<code>&lt;brief_title&gt;Assessment of Long-term Safety in Patients With Non-cancer-related Pain and Opioid-induced Constipation&lt;/brief_title&gt;</code></td>
<td>2</td>
<td>Incomplete input by ClinicalTrials.gov user</td>
</tr>
<tr>
<td>8</td>
<td>00924638</td>
<td>Results for the primary outcome</td>
<td><code>&lt;category&gt;&lt;sub_title&gt;Month 138&lt;br&gt;(N=79)&lt;/sub_title&gt;&lt;measurement_list&gt;&lt;measurement group_id=&quot;O1&quot; value=&quot;748.1&quot; lower_limit=&quot;614.3&quot; upper_limit=&quot;911.1&quot;/&gt;&lt;/measurement_list&gt;&lt;category&gt;&lt;sub_title&gt;Month 186&lt;br&gt;(N=98)&lt;/sub_title&gt;&lt;measurement_list&gt;&lt;measurement group_id=&quot;O1&quot; value=&quot;353.1&quot; lower_limit=&quot;284.0&quot; upper_limit=&quot;439.1&quot;/&gt;&lt;/measurement_list&gt;&lt;category&gt;&lt;sub_title&gt;Month 246&lt;br&gt;(N=85)&lt;/sub_title&gt;&lt;measurement_list&gt;&lt;measurement group_id=&quot;O1&quot; value=&quot;317.3&quot; lower_limit=&quot;247.4&quot; upper_limit=&quot;407.1&quot;/&gt;&lt;/measurement_list&gt;</code></td>
<td>6</td>
<td>Time series outcome measurements</td>
</tr>
<tr>
<td>9</td>
<td>00883246</td>
<td>Outcome</td>
<td><code>&lt;participant_flow&gt;…&lt;group group_id=&quot;P1&quot;&gt;&lt;title&gt;Atherectomy&lt;/title&gt;&lt;outcome_list&gt;…&lt;group group_id=&quot;O1&quot;&gt;&lt;title&gt;Claudicant Subgroup&lt;/title&gt;</code></td>
<td>6</td>
<td>Study arm changed without updating related information</td>
</tr>
</tbody>
</table>

Limitations

This study had several limitations. First, a small sample (n=100) of trial records was used to test rules for the automatic extraction of PICO elements from ClinicalTrials.gov. Future studies are needed to test generalizability of the handcrafted rules with a larger dataset. Second, the study was limited to ClinicalTrials.gov. There are other large clinical trial repositories that could be used to increase coverage, but different methods will be needed since there is not a standard format for trial registration across registries. Third, our findings are specific to high quality clinical trials published in high impact journals. Still, these are the trials that are most likely to be useful for clinical decision support. Last, we did not evaluate the usefulness and readability of extracted PICO elements. Future studies are needed to investigate the design and assess the usefulness and readability of extracted information for patient care decision-making. Another important step is to map the extracted information to standard controlled vocabularies whenever possible to increase utility of extracted PICO elements.

Conclusion

Our study found that the coverage of ClinicalTrials.gov for high quality trials published in high impact journals, although relatively low, has continuously increased over the last ten years. Our method, based on a set of rules, was able to accurately automatically extract the majority of PICO elements from a sample of 100 clinical trial records. The study suggests that using ClinicalTrials.gov for extracting key clinical trial information to support patient care...
decision-making is feasible. Future studies include assessing and tuning the algorithm on a larger sample size, expanding our work to other clinical trial repositories, developing structured information displays according to PICO elements, and assessing the usefulness of this type of information display for supporting patient care decisions.

Acknowledgement

This study was supported by grants number 1R01LM011416-01 and R00LM011389 from the National Library of Medicine (NLM).

References


Feasibility of Representing Data from Published Nursing Research Using the OMOP Common Data Model

Hyeoneui Kim, RN, MPH, PhD1, Jeeyae Choi, RN, MS, PhD2, Imho Jang3, Jimmy Quach4, Lucila Ohno-Machado, MD, MBA, PhD1

1Health System Department of Biomedical Informatics, UC San Diego, La Jolla, CA  
2College of Nursing, University of Wisconsin at Milwaukee, Milwaukee, WI  
3Division of Biological Sciences, UC San Diego, La Jolla, CA  
4Department of Bioengineering, UC San Diego, La Jolla, CA

Abstract

We explored the feasibility of representing nursing research data with the Observational Medical Outcomes Partners (OMOP) Common Data Model (CDM) to understand the challenges and opportunities in representing various types of health data not limited to diseases and drug treatments. We collected 1,431 unique data items from 256 nursing articles and mapped them to the OMOP CDM. A deeper level of mapping was explored by simulating 10 data search use cases. Although the majority of the data could be represented in the OMOP CDM, potential information loss was identified in contents related to patient reported outcomes, socio-economic information, and locally developed nursing intervention protocols. These areas will be further investigated in a follow up study. We will use lessons learned in this study to inform the metadata development efforts for data discovery.

Introduction and Background

Wide adoption of Electronic Health Record (EHR) systems and the development of secure cloud-based infrastructure to store large amounts of clinical and health research data have been opening new opportunities for data-driven discovery and validation of knowledge. Observational and pragmatic studies are gaining traction as alternatives to Randomized Controlled Trials (RCT) [1–3]. Although the highest quality of clinical evidence is still produced via RCTs [4], it is infeasible to perform an RCT for every clinical problem of interest for every possible patient situation. Pharmaco-surveillance [5,6], various comparative effectiveness studies [7,8] and cohort discoveries [9,10] are a few example areas that have benefited by the increasing availability of health data. However, these data also pose non-trivial challenges such as the handling of missing and/or noisy data and the integration of data generated from various sources, each with its own idiosyncratic representation. The Big Data to Knowledge (BD2K) initiative by NIH aims to address such challenges and to promote the utilization of large amounts of biomedical data (i.e., “big data”) by improving discoverability, accessibility, interoperability, and reusability of data and advanced data analytics [11,12].

bioCADDIE – indexing for big data discovery

bioCADDIE (biomedical and healthCAre Data Discovery Index Ecosystem) is a BD2K consortium dedicated to establishing a user friendly and robust means to discover and index biomedical data [13,14]. At its core are the bioCADDIE metadata, a minimum set of information about a dataset that needs to be made available to data seekers to facilitate search for data sets. The bioCADDIE metadata specification was developed by incorporating existing major metadata schemas in the biomedical domain (i.e., top-down) and high-priority dataset search use cases identified by the bioCADDIE leadership group and the user community. The version 1 metadata specification was released in August of 2015 and is currently under revision based on the feedback obtained from the bioCADDIE community [15]. Some of the referenced existing metadata and use cases reflect clinical and/or healthcare domains to a certain extent (mostly diagnoses, laboratory tests, and medications). However, overall, clinical and healthcare research domains are relatively under-represented in bioCADDIE metadata when compared to omics research domains. As a first step to augmenting the clinical and healthcare domains, the bioCADDIE team has started exploring the interoperability between the bioCADDIE metadata and the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), a widely accepted, standardized model for healthcare data [16].

OMOP CDM – a common data model for healthcare research data

OMOP was formed as a partnership between public and private sectors to establish the safe and effective use of observational healthcare data to study the effectiveness of medical products [17]. Observational Health Data
Sciences and Informatics (OHDSI) is an international collaboration that aims to promote data reusability of the observational healthcare data produced from various venues by achieving data interoperability among them [18]. The OMOP CDM is adopted and now continuously revised by OHDSI [19]. Patient Centered Outcomes Research network (PCORnet), a national network funded by PCORI, is another major effort that aims at utilizing large scale clinical and healthcare data research networks for patient-centered outcomes research. A CDM plays a critical role in achieving data interoperability [20–22]. PCORnet CDM [23] and OMOP have recently been cross mapped by the OHDSI research team [24].

Gaps in representing nursing data for big data science

Many studies explored the use of the OMOP CDM to conduct large scale drug surveillance studies as well as to integrate clinical data repositories [22,25–29]. However, few studies reported the experience on standardizing and integrating data generated from clinical or healthcare research with OMOP CDM, except for studies that involved drugs or medical devices.

The recent Precision Medicine Initiatives (PMI) calls for collecting and utilizing a full spectrum of biomedical data to improve patient care: from microbiology to physical exams, laboratory tests, behavioral and environmental exposures [30]. In addition, the National Academy of Medicine (NAM), formerly Institutes of Medicine (IOM), recommended the incorporation of social and behavioral domains of data in EHRs in such a way that would support the reuse of those data in patient care and research [31]. Therefore, integrating different types of data in a way that supports secondary analyses has become critical. Nursing research targets a wide range of nursing interventions and problems, which often may not involve specific drugs, medical devices, or medical diagnoses. It often includes topics such as patients’ functional status, education, satisfaction toward the provided care and quality of life. This might indicate that nursing and nursing research data present a unique opportunity to understand the challenges and special considerations in integrating data generated in variety of health domains. However, few studies investigated the feasibility of representing various types of nursing and nursing research data with a CDM except for a very specific nursing domain such as pressure ulcer risk assessment and prevention [32,33].

The purpose of this study was to test the feasibility of representing nursing research data with the OMOP CDM using data reported in published articles and subsequently indexing them with the bioCADDIE tools. Specifically, we evaluated (1) the content coverage of the OMOP CDM (version 5) toward the data content reported in 256 published nursing research articles and (2) the feasibility of conducting data searches using the use cases extracted from the 10 randomly selected research articles.

Materials and Methods

Data preparation

In a prior separate study, we reviewed 192 articles on clinical trials published in nursing journals between 2006-2015. From this review, we annotated the articles with a list of metadata such as study types, target research problems, key interventions, comparison groups, sample demographics, data items used/generated, data sources, funding supports, data access information, and locations and settings of the studies. To augment this collection with observational studies we conducted additional literature review. Using PubMed and its structured search filters, we retrieved articles classified as Observational Studies published in Nursing Journals for the past 3 years (i.e., 2013 ~ February of 2016). This search yielded 203 articles. We randomly selected about half of the retrieved articles (N=98). We then read the full text articles and annotated each article with the same set of metadata.

During the full text review we excluded 34 articles for which full text papers were not easily obtainable (N=3) or the articles that did not investigate patient problems from nursing perspectives thus consequently no patient outcomes were reported (N=31). For example, behavioral issues among nursing students [34], assessment scale development [35], observing hand washing behavior among nurses without reporting associated patient safety outcomes [36] and nursing management topics such as nursing shifts and workload estimation [37,38] were excluded from this review. With this additional review, we generated metadata for 64 articles on nursing observational studies. Therefore, combining the new articles with the previously tagged ones, we included a total of 256 articles to this study. The article selection process is summarized in Figure 1. We also classified the data items reported in 256 articles based on the modes of data collection – i.e., routinely collected clinical data, Patient Reported Outcomes (PRO), and data collected through mobile devices.
Mapping to the OMOP CDM

A total of 2,438 data items were extracted from the 256 articles. After removing duplicate entries we classified 1,431 unique data items into concept classes of the OMOP CDM to establish a high (i.e., general) level of mapping. This mapping was largely guided by the definition and usage description provided for each concept class in OMOP CDM. However, we also checked the attributes of concept classes to ensure a given data item had a relevant attribute that could be used to hold the information about the data item. For example, one might expect that various socio-economic, and demographic items such as marital status and education level would belong to the PERSON class as done in many healthcare databases. However OMOP’s PERSON class does not provide a place to hold those items as it was designed to hold permanent information about a person such as birth year and race. Two of the authors (HK, JC) collaboratively conducted this high level mapping. We also aimed at identifying data items that did not fit into any of the OMOP CDM classes during this high level mapping activity.

Next we conducted a small scale, deeper level mapping and investigated to what extent the semantics of the content items presented in the selected articles were represented in the OMOP CDM. We fully utilized the attributes and the constraints associated with the concept classes for this mapping. We first created 10 data search use cases from 10 articles randomly selected from the 256 articles pool by asking ourselves “how do you query a clinical data repository to replicate the study reported in this article?” These use cases incorporated the eligibility criteria of the studies and thus provided an opportunity to evaluate additional content items not reflected in the data items at the deeper level (i.e., value level). Next we simulated the dataset search use cases by converting them into a data query form similar to the SQL code style (i.e., pseudo SQL codes). We used the mapping established between the OMOP CDM and the data items as a database schema of a hypothetical data repository. A use case and its data query form are presented as an example in Figure 2.

To simplify the conversion we omitted certain details and employed a few of our own conventions: we did not specify the primary key and foreign key relations to simplify the conversion. We used standardized concept id
attributes (e.g., gender_concept_id, observation_concept_id, etc.) to represent data item names and non-numeric data item values. Instead of specifying mapped standardized concept codes, we put data item or value names in brackets to imply the values are the concept codes of the concept names presented in the brackets. For example, the code observation_concept_id [ ] = [marital status] in Figure 2 implies that the observation_concept_id attribute takes the concept code of an OMOP-recognized standardized vocabulary of choice, for instance “125680007” when the Systematized Nomenclature of Medicine- Clinical Term (SNOMED-CT) is used to encode values.

Figure 2 Data search use case and data query form translation

Results
The majority of the data items identified from the 256 articles were mapped to the concept classes in the OMOP CDM, except for 16 data items. A few examples of unmapped items were the socio-economic and/or demographic information of health care providers (e.g., “number of years working as a nurse”) and the items that potentially embedded multiple content items, each of which would need to have been mapped to different concept classes. For example, “operating room nurses’ task list” and “advice provided by healthcare provider” might need to be mapped to different classes depending on what values they take.

MEASUREMENT was the concept class that most data items were related to, followed by OBSERVATION. The data items mapped to each concept class in the OMOP CDM is summarized in examples in Table 1.

PRO items collected using a questionnaire (both standardized or custom developed) were reported in 82 articles and
data generated through mobile devices were reported in 2 articles.

Our attempt to simulate data search from the use cases was not completely successful in all 10 cases. We observed potential information loss in 2 cases where (1) the study outcomes were assessed by comparing dates and times of a surgical procedure and (2) a device study that required dates and times of certain clinical findings. The PROCEDURE and DEVICE classes of the OMOP CDM stored only date information, thus precise comparison was not deemed possible. The metadata annotated for the 256 articles, the full results of CDM mapping, and data search use case simulations are available to download from https://idash-data.ucsd.edu/download/folder/4825/AMIA2016.zip.

<table>
<thead>
<tr>
<th>Mapped OMOP CDM Class</th>
<th>Count (N=1,431)</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement</td>
<td>529</td>
<td>Charlson comorbidity index score, Ejection fraction, Vital signs</td>
</tr>
<tr>
<td>Observation</td>
<td>319</td>
<td>Treatment history, Smoking status, Marital status, Catheter leakage</td>
</tr>
<tr>
<td>Condition_Occurrence</td>
<td>223</td>
<td>Problem list, Admitting diagnoses</td>
</tr>
<tr>
<td>Procedure_Occurrence</td>
<td>109</td>
<td>Renal transplant, Chemotherapy, Detoxification protocol, Relaxation</td>
</tr>
<tr>
<td>Person</td>
<td>94</td>
<td>Age, Race, Gender, Ethnicity</td>
</tr>
<tr>
<td>Drug_Exposure</td>
<td>48</td>
<td>Opioid dose, Type of opioid used, Pain medication</td>
</tr>
<tr>
<td>Device_Exposure</td>
<td>25</td>
<td>Mechanical ventilator use, Traction device type, CPAP use</td>
</tr>
<tr>
<td>Visit_Occurrence</td>
<td>27</td>
<td>Admission date, Discharge date</td>
</tr>
<tr>
<td>Care_Site</td>
<td>15</td>
<td>Type of hospital, Admitting service</td>
</tr>
<tr>
<td>Provider</td>
<td>3</td>
<td>APN specialty, Service provider type</td>
</tr>
<tr>
<td>Payer_Plan_period</td>
<td>6</td>
<td>Payer, Health insurance type, Insurance status</td>
</tr>
<tr>
<td>Visit_Cost</td>
<td>5</td>
<td>Average cost per patient per patient day, Average cost per patient</td>
</tr>
<tr>
<td>Drug_Era</td>
<td>4</td>
<td>Period of hormone therapy, Period of chemotherapy</td>
</tr>
<tr>
<td>Death</td>
<td>4</td>
<td>Hospital death, Death</td>
</tr>
<tr>
<td>Procedure_Cost</td>
<td>3</td>
<td>Wound dressing cost, Overhead</td>
</tr>
<tr>
<td>UNMAPPED</td>
<td>17</td>
<td>Influence of ward condition, Counselor’s age, Hygiene compliance</td>
</tr>
</tbody>
</table>

Discussion

We attempted to represent the data contents reported in 256 nursing articles on health studies using the OMOP CDM version 5. Our goal was to start understanding the scope of the representation supported by the OMOP CDM, a major information model standard, with regards to content areas not related to conventional comparative effectiveness research, using the data from nursing research. It is promising that the majority of the clinical contents used in nursing research were well represented in the OMOP CDM, especially through the Standardized Clinical Data domain. However, it is our conclusion that the OBSERVATION and the MEASUREMENT classes had to be somewhat abused in this mapping as a place to hold many non-clinical and/or unusual clinical findings and assessment items.

Representing socio-economic and demographic information of patients, family members, and healthcare providers

PERSON and PROVIDER are two classes of the OMOP CDM that hold information on human subjects in the healthcare domain. The former is dedicated to describing patients, and the latter describes healthcare providers. Family dynamics is expected to have a significant impact on patient outcomes therefore it is an important research topic in nursing. Many articles reviewed in our study reported information about family members, especially those who play the role of a main care taker. We mapped those items into the OBSERVATION class, although this class was designed to capture the clinical facts not represented in other concept classes in the OMOP CDM. Similarly, many socio-economic items (e.g., marital status, education level, employment status, etc.) about patients had to be mapped to the OBSERVATION class since attributes in the PERSON class are limited to key “unchangeable” demographics such as age, gender, race, and ethnicity. Similar challenges were observed with the socio-
demographic information about healthcare provider. We did not map these items to OBSERVATION, as the latter is to capture “any clinical facts about a patient obtained in the context of examination, questioning, or a procedure”[17]. These additional socio-demographic items on healthcare providers seem to be potential areas for expansion to fully support observational studies in nursing.

Representing patient reported outcomes

A third of the articles reviewed in this study reported assessment of patient status using standardized and/or custom-built assessment scales and survey questionnaires. Some of these assessment tools are completed by patients and some are by healthcare providers. Differentiating the source of information is important as it provides additional context to consider when studying patient outcomes. Of note, the CONDITION table of the PCORnet CDM captures this information through the condition_source attribute, whose structured value set contains patient reported medical history. It might be possible to infer the source of information for the outcomes measured with standardized scales that are widely used and well documented, although this practice still increases the risk for information loss by requiring additional steps for obtaining and associating the information. However, the source of information for outcomes measured with custom-built survey questionnaire is still highly likely be lost.

Representing nursing specific intervention concepts

Many articles we reviewed reported the effectiveness of specific nursing interventions, which often include new care protocols and complementary therapies. As recognized by the leaders in nursing informatics as a high priority task in nursing informatics, having nursing specific data sufficiently represented in “big data” sets is important to promote the incorporation of nursing’s holistic approaches to patient care into data-driven knowledge generation [39,40]. Despite the relative small scale of concept mapping for nursing interventions pursued in this study, we found that the majority of complementary therapies are covered by SNOMED-CT. However, locally developed specific care protocols are challenging to represent, since they are built from a collection of nursing actions. Consequently, a care protocol name used to describe an intervention concept might be considered less informative than the listing of nursing actions. Although it seems somewhat convoluted, this challenge can be addressed by establishing a member relation between a set of nursing interventions and a specific care protocol through the FACT_RELATIONSHIP table. We will continue investigating this issue in future studies.

Nursing problems cover wide areas of health domains: from disease-oriented matters to emotional wellbeing. Nursing research thus also includes concepts there are not frequently used in other clinical studies. However, we found that the OMOP CDM provides sufficient representation capability, even for unusual nursing problem concepts through less restrictive concept classes like OBSERVATION and MEASUREMENT, along with CONDITION_OCCURRENCE.

Limitations and future directions

This study used data items reported in a small sample of nursing research articles. Therefore, its findings may not be generalized to the entire content domains of nursing research data. In addition, the OMOP CDM strives for a high level of standardization and hence all key concepts in the CDM classes are required to be encoded with a standardized vocabulary. This means that the class and attribute levels of association does not always result in successful representation of nursing research concepts, as information loss can occur due to the lack of a standardized concept code. Using the OMOP CDM as a hypothetical database model to simulate data search use cases is another limitation. Information loss can happen during the transformation of a local database model to the OMOP CDM, and our simulation approach might not have been robust enough to capture this type of loss.

Recognizing these limitations, along with potential over-use of certain concept classes as previously described, we plan to continue this work by expanding the scope and revisiting identifiably challenging cases. We also plan to seek collaboration with OHDSI to discuss and substantiate lessons learned in this study.

Conclusion

Nursing research targets wide ranges of topic areas that are not limited to diseases or medical treatments. Therefore, investigating the standardized representation of the data contents in published nursing research provides a unique opportunity to understand challenges and opportunities in representing various types of health data so that they can be found and reused. In this study, we explored the feasibility of representing nursing research data with the OMOP CDM, using the data items reported in 250 nursing articles. The OMOP CDM provided a good representation for the majority of data items but we also observed potential gaps that might lead to information loss. These gaps will be
further investigated in a follow-up study. Lessons learned in this study will help inform metadata development efforts for the Data Discovery Index.

Acknowledgment

This project was supported in part by the grant 1U24AI117966 (NIH/BD2K).

References


13. bioCADDIE | biomedical and healthCAre Data Discovery and Indexing Ecosystem [Internet]. [cited 2015 May 31]. Available from: https://biocaddie.org/


18. OHDSI | Observational Health Data Sciences and Informatics [Internet]. [cited 2015 Nov 29]. Available from: http://www.ohdsi.org/


Analysis of Healthcare Cost and Utilization in the First Two Years of the Medicare Shared Savings Program Using Big Data from the CMS Enclave

Fabricio S. P. Kury, MD1, Seo H. Baik, PhD1, Clement J. McDonald, MD1
1Lister Hill National Center for Biomedical Communications, National Library of Medicine, U.S. National Institutes of Health, Bethesda, Maryland, USA

Abstract

The Medicare Shared Savings Program (MSSP) is the larger of the first two Accountable Care Organization (ACO) programs by the Centers for Medicare and Medicaid Services (CMS). In this study we assessed healthcare cost and utilization of 1.71 million Medicare beneficiaries assigned to the 333 MSSP ACOs in the calendar years of 2013 and 2014, in comparison to years 2010 and 2011, using the official CMS data. We employed doubly robust estimation (propensity score weighting followed by generalized linear regression) to adjust the analyses to beneficiary personal traits, history of chronic conditions, previous healthcare utilization, ACO administrative region, and ZIP code socioeconomic factors. In comparison to the care delivered to the control cohort of 17.7 million non-ACO beneficiaries, we found that the care patterns for ACO beneficiaries shifted away from some costly types of care, but at the expense of increased utilization of other types, increased imaging and testing expenditures, and increased medication use, with overall net greater increase in cost instead of smaller increase.

Introduction

The Medicare Shared Savings Program (MSSP) was launched by the Centers for Medicare and Medicaid Services in 2012 as one of the first two Accountable Care Organization (ACO) programs as envisioned by the Patient Protection and Affordable Care Act (PPACA), which seek to coordinate and improve healthcare delivery by means of value-based reimbursement models. Considerable work has been done to study ACOs, including assessments of previously existing pilot programs and state-specific value-based reimbursement models1-7, early studies of expected implications and qualitative analyses of impacts8-10, studies focused on specific medical specialties2,3,10-16, and more recent data-based reviews of the Pioneer ACO program, also launched in 201217,18.

In 2015, CMS made available, through their privacy-protecting Chronic Conditions Data Warehouse19 (CCW), the complete list of all beneficiaries assigned to and all providers enrolled in MSSP ACOs until December 31st, 2013. In 2016, the continuation of this dataset was released which extended data availability until the end of 2014. Combined, these two datasets cover all 333 ACOs that joined the MSSP between April 1st, 2012 and January 1st, 2014, and all Medicare beneficiaries assigned to them. The retrospective and detailed nature of the algorithm that CMS uses to assign beneficiaries to ACOs requires knowing the proportion of the total care received by each beneficiary that was delivered by ACO-enrolled providers, as well as the specialty of the physicians involved20, and this information is difficult to infer without the authoritative datasets that CMS released in 2015 and 2016. As of July, 2016, there is a single publication32 analyzing the CMS MSSP data released in 2015, but no published analysis using both 2015 and 2016 datasets to assess the effect of the MSSP ACOs on resource utilization and cost. Here we report such an analysis. For simplicity, starting from this sentence the term “ACO” will always mean only MSSP ACO, never Pioneer ACO, except at the Discussion section of the paper.

Methods

Reproducible research

As a commitment to the transparency and reproducibility of our work, we voluntarily release for free download the complete source code used in this paper, under an Attribution-NonCommercial-ShareAlike 4.0 International29 license, in an online GitHub repository at the link http://github.com/fabkury/cisa. It is important to remind here that for re-executing the code an interested researcher needs his/her own Data Use Agreement with CMS, by which one can receive access to the CMS claims files. The GitHub repository provides solely the source code and the full set of results produced by that source code.
Study design

We compared cost and utilization in the care provided to the beneficiaries assigned to ACOs with those not assigned using a difference-in-differences analysis of the trends in each group between before and after the start of the MSSP. The before period (baseline period) we defined as the combination of years 2010 and 2011, while the after period (outcome period) was the combination of 2013 and 2014. We ruled to completely ignore year 2012 as means of minimizing noise, upon recognizing that several of the most important variables we desired for analysis were available only at year-level granularity, while some ACOs had their agreement start date in April 1st, 2012 and some in July 1st, 2012. All ACOs from 2013 dataset (released in 2015) had their start date on or before January 1st, 2013, and those on the 2014 dataset (released in 2016) had their start date on or before January 1st, 2014. We additionally ruled to ignore any beneficiary that changed his or her U.S. state of residence at any moment between 2010 and 2014, as this could impart noise due to different healthcare pricing levels at different states.

Inpatient, Outpatient, Skilled Nursing Facility (SNF) and Physician Part B (Carrier) claims were aggregated by these four types as they are available in their separate CMS Research Identifiable Files (RIFs). Components of care such as Intensive Care Unit days, hospice days, imaging and laboratory events and expenditures, Home Health visits, among others, were aggregated separately as available in the CCW Master Beneficiary Summary and MedPAR files, which identify these items using revenue codes.

Beneficiary eligibility for this study

To be sure that we had complete claims data needed for these analyses, we only included beneficiaries who were fully enrolled in Medicare Parts A and B during the entire years of 2010, 2011, 2013 and 2014, and had zero months of Health Maintenance Organization (HMO, Medicare Part C) enrollment in those four years. This also means that any beneficiary either not enrolled in Medicare since January 1st, 2010, or not alive until December 31st, 2014, was excluded. For defining the cohort of ACO-assigned beneficiaries, we further only considered those with ACO assignment during both complete calendar years of 2013 and 2014. For the cohort of non-ACO beneficiaries, we considered only those never assigned to an ACO at any moment.

Figure 1. Beneficiary eligibility

<table>
<thead>
<tr>
<th>Medicare beneficiaries extant in all years 2010, 2011, 2013 and 2014.</th>
<th>41,873,231</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Medicare Parts A and B enrollment in all four years.</td>
<td>36,919,919</td>
</tr>
<tr>
<td>Zero HMO (Medicare Part C) enrollment.</td>
<td>23,432,694</td>
</tr>
<tr>
<td>No change of U.S. state of residence during study period.</td>
<td>22,422,157</td>
</tr>
<tr>
<td>Assigned to ACO during entire years 2013 and 2014.</td>
<td>1,708,888</td>
</tr>
</tbody>
</table>

Statistical analysis

We first used the propensity score approach to mitigate differences in beneficiary-level characteristics between those enrolled in ACO and those not. We employed a binary logistic regression model to calculate the predicted probability of enrolling in an ACO using 66 covariates obtained from the baseline period (years 2010 and 2011). Covariates used in the propensity score included beneficiary-level characteristics such as age (at the end of 2011), gender, race, presence of 26 CMS predefined priority chronic conditions (by the end of 2011), region of residence as defined by
the ACO regional offices, reason for Medicare eligibility, and several separate counts and costs of drug and non-drug care received. Each possible region of residence, chronic health condition, race, reason for Medicare eligibility, as well as gender, was a binary flag. The remaining covariates were numerical values.

We then used a generalized linear regression model with the propensity score weight to compare outcomes of interest between those enrolled in ACOs and those not. The propensity score weight was defined as the inverse of the predicted probability of enrolling in an ACO. Individuals with a high predicted probability of enrolling in ACOs, who actually enrolled in ACOs, receive a lower weight compared with individuals with a low probability of enrolling in ACOs. In this regression model, outcomes of interest from the ACO and the non-ACO groups were compared not only weighted by the propensity score, but also controlled again for beneficiary-level and also beneficiary ZIP code-level characteristics, forming a doubly robust estimator. The generalized linear regression model included 48 covariates, of which 45 were also included in the propensity score. In total, the statistical analyses adjusted for 69 covariates. We conducted all statistical analyses using SAS software, version 9.4 (SAS Institute Inc., Cary, NC), inside the CMS Virtual Research Data Center (CMS “Enclave”).

For economy of space, the covariates included in the propensity score and in the linear regression model are specified together with their balances before and after the propensity score weighting in Table 1 at the Results section. All 26 chronic health conditions were included in both propensity score and linear regression, but were omitted from the table in this paper for economy of space – they are available in the full set of results for download at the online repository.

Outcome measures

We considered an extensive list of 151 outcome measures, as well as separate measurements of number of claims and total cost of filled drug prescriptions aggregated by 493 different ATC-4 classes that matched the NDCs present in the Medicare Part D claims of the beneficiaries included in the study, totaling 1,137 comparisons between the ACO cohort and the Non-ACO cohort. Due to the obvious lack of space, we opted to include and discuss in this paper only the differences that were most insightful to our opinion. The selected measures are depicted in Tables 2, 3 and 4 at the Results section. The complete list of outcome measures can be downloaded, along with our complete source code, at the online repository.

Part D drug claims and cost measures

The measurements of drug claims and costs aggregated by their ATC-4 classes (which were mapped from the drugs’ NDC codes) are not presented in this paper. They are available only for download at the online repository. We made this decision for three reasons. First, we preferred to prioritize other content in our limited space in this paper. Second, this specific part of the analysis relies solely on a 10% purely random sample of the beneficiaries, which is the sub-cohort from whom we have Part D claims as per our Data Use Agreement with CMS, while all other parts of this paper enjoy access to the totality of Medicare beneficiaries that met our eligibility criteria for this study. Third, the costs of Part D claims do not count towards the performance of a MSSP ACO as per the program’s definition, therefore the ACOs have no direct incentive to save on expenses that are billed to Part D. None of these three reasons are the case of drugs administrated inside health care facilities and billed to Medicare Part B. The costs of all drugs billed to Part B for all beneficiaries are included in the drug usage measurements in Table 3. Table 4 contains, separately, the total sum of costs of drugs billed to both Medicare Parts B and D and covers the totality of beneficiaries (not 10%).

Exclusion of outlier claims

Because our analyses processed hundreds of millions of individual claims from millions of beneficiaries, we were concerned about potential outlier values among the claims, e.g. data entry errors or faulty data for any reason. In order to mitigate the potential impact of unknown outliers on our calculations of averages, we implemented a mechanism for ignoring extreme outlier claims. Any claim whose cost ranked at the top 0.01\textsuperscript{th} percentile (0.01%, not to be confused with top 1\textsuperscript{st} percentile) of its data source (i.e. each CMS file available in the VRDC) was deemed a potential outlier and therefore ignored.

Data Sources

This study was performed using the CMS claims data available to our group inside the CMS Virtual Research Data Center via our Data Use Agreement #26250. The claims files used were full-size (100% of the Medicare beneficiaries) Inpatient RIF, Outpatient RIF, Skilled Nursing Facility RIF, Carrier RIF, MedPAR, and Master Beneficiary Summary File segments A/B, Chronic Conditions, and Cost & Use. Socioeconomic variables for each ZIP Code Tabulation Area (ZCTA) were retrieved from the FactFinder tool at Census.gov. The mapping of
beneficiary ZIP code to ZCTA was performed using the crosswalk table produced by the Robert Graham Center upon funding by the Health Resources and Services Administration. The mapping of Medicare Part D drugs from their NDC codes to ATC-4 classes was performed by querying the RxNorm API web service using an R language script, which is also available for download in the online repository.

Results

After applying all restraints described in the Methods section, we had a cohort of 1,708,888 ACO-assigned beneficiaries and a control cohort of 17,741,210 Non-ACO beneficiaries. The outcomes included 151 measurements plus 986 drug class utilization and cost measurements, so the total number of outcome comparisons between the ACO cohort and the non-ACO cohort was 1,137. Due to the very large size of the studied cohorts, the p-values were very small, although in many cases the identified differences was also small if not frankly negligible. In the great majority of the cases, the resulting p-value was smaller than 10^{-6}, so we thought there was little need to perform corrections for multiple-testing, as it was unlikely that we would arrive at a non-negligible probability of having committed a statistical Type I (“false positive”) error. Unless otherwise indicated, a p-value below 10^{-6} is to be assumed for all comparisons included in this paper, except for those in Table 1. The p-values of all comparisons (including those of Table 1) can be found, rounded to 9 decimal places, in complete set of results available for download at the online repository. All monetary values are expressed in U.S. dollars.

Covariate balancing by the propensity score

Table 1 depicts all variables included in the propensity score (PS) – except for the 26 CCW chronic conditions, here omitted for economy of space, which were also included – and their balance before and after weighting. Due to the very large number of beneficiaries in both ACO and non-ACO cohorts, even despite the small differences in the averaged values after weighting, all but 9 p-values when comparing the two groups post-weighting were below 10^{-6}.

Table 1. Cohort characteristics and covariates used in the statistical analyses, along with average values pre and post propensity score weighting, except for the 26 chronic health conditions. Numbers are average per beneficiary in years 2010 and 2011 ± standard error. The binary flags are indicated by the absolute count followed by the percentage of the total that the count represents.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Non-ACO</th>
<th>ACO</th>
<th>Non-ACO</th>
<th>ACO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beneficiary age at the end of 2011</td>
<td>70.859 ± 12.334</td>
<td>72.723 ± 10.817</td>
<td>71.022 ± 13.015</td>
<td>71.057 ± 39.851</td>
</tr>
<tr>
<td>Female gender</td>
<td>9,604,042 (54.27%)</td>
<td>998,475 (58.51%)</td>
<td>9,671,501 (54.65%)</td>
<td>954,972 (55.96%)</td>
</tr>
<tr>
<td>White</td>
<td>14,969,286 (84.59%)</td>
<td>1,499,954 (87.90%)</td>
<td>15,021,505 (84.88%)</td>
<td>1,455,965 (85.32%)</td>
</tr>
<tr>
<td>Black</td>
<td>1,693,781 (9.57%)</td>
<td>126,223 (7.40%)</td>
<td>1,660,070 (9.38%)</td>
<td>160,380 (9.40%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>340,354 (1.92%)</td>
<td>20,635 (1.21%)</td>
<td>329,642 (1.86%)</td>
<td>31,978 (1.87%)</td>
</tr>
<tr>
<td>Asian</td>
<td>275,298 (1.56%)</td>
<td>32,341 (1.90%)</td>
<td>279,704 (1.58%)</td>
<td>23,114 (1.36%)</td>
</tr>
<tr>
<td>Other ethnicity</td>
<td>418,412 (2.36%)</td>
<td>27,269 (1.60%)</td>
<td>406,210 (2.30%)</td>
<td>34,986 (2.05%)</td>
</tr>
<tr>
<td>ACO region 1: MA, VT, NH, ME, RI, CT</td>
<td>924,786 (5.23%)</td>
<td>220,998 (12.95%)</td>
<td>1,046,749 (5.92%)</td>
<td>103,225 (6.05%)</td>
</tr>
<tr>
<td>ACO region 2: NY, NJ, PR, USVI</td>
<td>1,617,822 (9.14%)</td>
<td>229,682 (13.46%)</td>
<td>1,685,348 (9.52%)</td>
<td>157,553 (9.23%)</td>
</tr>
<tr>
<td>ACO region 3: PA, DE, WV, VA, DC, MD</td>
<td>1,922,113 (10.86%)</td>
<td>117,078 (6.86%)</td>
<td>1,859,328 (10.51%)</td>
<td>180,397 (10.57%)</td>
</tr>
<tr>
<td>ACO region 4: GA, KY, TN, NC, SC, AL, MS, FL §</td>
<td>3,885,810 (21.96%)</td>
<td>365,902 (21.44%)</td>
<td>3,877,755 (21.91%)</td>
<td>373,615 (21.90%)</td>
</tr>
<tr>
<td>ACO region 5: IL, MN, WI, MI, IN, OH</td>
<td>2,859,615 (16.16%)</td>
<td>393,841 (23.08%)</td>
<td>2,967,123 (16.77%)</td>
<td>288,871 (16.93%)</td>
</tr>
<tr>
<td>ACO region 6: TX, NM, OK, AR, LA §</td>
<td>2,189,383 (12.37%)</td>
<td>88,179 (5.17%)</td>
<td>2,076,514 (11.73%)</td>
<td>200,563 (11.75%)</td>
</tr>
<tr>
<td>ACO region 7: MO, NE, IA, KS</td>
<td>960,840 (5.43%)</td>
<td>127,170 (7.45%)</td>
<td>992,413 (5.61%)</td>
<td>98,240 (5.76%)</td>
</tr>
<tr>
<td>ACO region 8: CO, MT, ND, WY, SD, UT §</td>
<td>575,102 (3.25%)</td>
<td>19,051 (1.12%)</td>
<td>541,703 (3.06%)</td>
<td>52,351 (3.07%)</td>
</tr>
<tr>
<td>ACO region 9: CA, NV, AZ, HI, GU, AS, MP</td>
<td>2,020,863 (11.42%)</td>
<td>106,485 (6.24%)</td>
<td>1,940,036 (10.96%)</td>
<td>183,422 (10.75%)</td>
</tr>
<tr>
<td>ACO region 10: WA, OR, ID, AK §</td>
<td>740,797 (4.17%)</td>
<td>38,036 (2.23%)</td>
<td>710,162 (4.01%)</td>
<td>68,206 (4.00%)</td>
</tr>
<tr>
<td>Medicare-elig. due to age &gt;= 65</td>
<td>14,063,975 (79.47%)</td>
<td>1,461,251 (85.63%)</td>
<td>14,159,233 (80.01%)</td>
<td>1,368,480 (80.20%)</td>
</tr>
<tr>
<td>Medicare-elig. due to disability</td>
<td>3,545,624 (20.04%)</td>
<td>238,818 (14.00%)</td>
<td>3,450,985 (19.50%)</td>
<td>328,552 (19.25%)</td>
</tr>
<tr>
<td>Medicare-elig. due to end-stage renal disease</td>
<td>39,273 (0.22%)</td>
<td>2,479 (0.16%)</td>
<td>39,298 (0.22%)</td>
<td>4,069 (0.24%)</td>
</tr>
<tr>
<td>Medicare-elig. due to disab. and renal disease</td>
<td>48,259 (0.27%)</td>
<td>3,874 (0.23%)</td>
<td>47,616 (0.27%)</td>
<td>5,322 (0.31%)</td>
</tr>
<tr>
<td>ZIP code pop. below Federal Poverty Level</td>
<td>0.154 ± 0.095</td>
<td>0.133 ± 0.089</td>
<td>0.153 ± 0.099</td>
<td>0.141 ± 0.303</td>
</tr>
<tr>
<td>ZIP code pop. &gt;=25 years with High School degree</td>
<td>0.862 ± 0.089</td>
<td>0.88 ± 0.081</td>
<td>0.863 ± 0.092</td>
<td>0.871 ± 0.298</td>
</tr>
<tr>
<td>Number of chronic conditions at end of 2011</td>
<td>5.475 ± 3.752</td>
<td>6.581 ± 3.293</td>
<td>5.575 ± 3.925</td>
<td>5.88 ± 11.44</td>
</tr>
<tr>
<td>Part D Fill Count</td>
<td>47,822 ± 73.194</td>
<td>55,033 ± 75.034</td>
<td>48,492 ± 77.121</td>
<td>51,977 ± 243.641</td>
</tr>
<tr>
<td>Part B Physician Events</td>
<td>13,132 ± 14.271</td>
<td>17.525 ± 14.611</td>
<td>13,569 ± 15.9</td>
<td>14,916 ± 42.823</td>
</tr>
</tbody>
</table>
Out of the 69 covariates considered, 58 had a p-value below 10^{-6} even after the propensity score weighting. Despite the low p-values, as one could expect from an analysis totaling 19.45 million individuals, the propensity score was effective in balancing their means between the ACO and non-ACO cohorts. The full list of covariates, including the 26 chronic health conditions and all p-values, can be downloaded at the GitHub repository.

Comparison of overall claims by type of healthcare facility

Table 2 shows that the average number of claims per beneficiary, as well as average cost per beneficiary, increased more in the ACO cohort than in the non-ACO in outpatient and physician Part B (carrier) settings, while at inpatient and skilled nursing facility settings it was the opposite. The numbers under the Non-ACO and ACO columns in Tables 2, 3 and 4 are the difference between the average per beneficiary in the baseline and outcome periods, i.e. average per beneficiary at outcome period minus average per beneficiary at baseline period, along with the standard error of the average. The Difference column is defined as Non-ACO value minus ACO value, while the Percent difference column is [Non-ACO value minus ACO value] divided by unsigned Non-ACO value. Therefore, in those three tables, negative differences (as well as negative percent differences) indicate that the average per beneficiary increased more in the ACOs cohort than in the non-ACO cohort from the baseline (before) period to the after (outcome) period.

Table 2. Claims aggregated by type of healthcare facility. Numbers are differences in average per beneficiary between the baseline and outcome periods.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Non-ACO</th>
<th>ACO</th>
<th>Difference</th>
<th>Percent diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier (physician Part B) claims</td>
<td>5.54 ± 0.0129</td>
<td>3.985 ± 0.0133</td>
<td>-1.5574</td>
<td>-28.27%</td>
</tr>
<tr>
<td>Carrier (physician Part B) total charges ($)</td>
<td>679.30 ± 2.22</td>
<td>1033.57 ± 2.28</td>
<td>-354.27</td>
<td>-52.15%</td>
</tr>
<tr>
<td>Outpatient claims</td>
<td>1.379 ± 0.0033</td>
<td>1.534 ± 0.0034</td>
<td>-0.1554</td>
<td>-11.20%</td>
</tr>
<tr>
<td>Outpatient total charges ($)</td>
<td>10,077 ± 28.41</td>
<td>11,036.73 ± 29.22</td>
<td>-959</td>
<td>-9.52%</td>
</tr>
<tr>
<td>Inpatient total charges ($)</td>
<td>6,270.27 ± 26.359</td>
<td>5,610.71 ± 27.12</td>
<td>659.55</td>
<td>10.52%</td>
</tr>
<tr>
<td>Inpatient claims</td>
<td>0.08 ± 0.0004</td>
<td>0.071 ± 0.0004</td>
<td>0.0089</td>
<td>11.10%</td>
</tr>
<tr>
<td>Skilled Nursing Facility total charges ($)</td>
<td>1,287.48 ± 4.28</td>
<td>886.70 ± 4.41</td>
<td>400.79</td>
<td>31.13%</td>
</tr>
<tr>
<td>Skilled Nursing Facility claims</td>
<td>0.105 ± 0.0004</td>
<td>0.068 ± 0.0004</td>
<td>0.0366</td>
<td>34.86%</td>
</tr>
</tbody>
</table>

Comparison of components of care at inpatient and skilled nursing facility settings

Table 3 depicts some of the outcome measures available from the MedPAR file, which covers inpatient and skilled nursing facility (SNF) claims. It seemingly shows a trend of health care shifting away from expensive types of care such as stays in heart transplant and myocardial infarction specialized Coronary Care Units (CCUs), at the expense of increased use of intermediate CCUs and ICUs and increased expenditures with therapeutic Radiology. In these two care settings (inpatient and SNF), diagnostic radiology claims, laboratory charges and ambulatory surgical care claims
increased faster in the Non-ACO cohort. The overall net effect here was positive for the ACO cohort, i.e. smaller increase in total cost.

Table 3. Components of care at inpatient and skilled nursing facility settings. Numbers are differences in average per beneficiary between the baseline and outcome periods.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Non-ACO</th>
<th>ACO</th>
<th>Difference</th>
<th>Percent diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary care unit, pulmonary claims †</td>
<td>0.000037 ± 0.000006</td>
<td>0.000056 ± 0.000006</td>
<td>-0.0000195</td>
<td>-52.53%</td>
</tr>
<tr>
<td>Coronary care unit, intermediate claims</td>
<td>0.006172 ± 0.000098</td>
<td>0.008709 ± 0.0001</td>
<td>-0.0025371</td>
<td>-41.11%</td>
</tr>
<tr>
<td>ICU, intermediate IOU claims</td>
<td>0.017504 ± 0.00138</td>
<td>0.022418 ± 0.00142</td>
<td>-0.0049134</td>
<td>-28.07%</td>
</tr>
<tr>
<td>Radiology therapeutic claims</td>
<td>0.000498 ± 0.00002</td>
<td>0.000622 ± 0.00002</td>
<td>-0.0001236</td>
<td>-24.80%</td>
</tr>
<tr>
<td>Coronary Care Unit days ‡</td>
<td>0.037944 ± 0.000546</td>
<td>0.039248 ± 0.000562</td>
<td>-0.0013036</td>
<td>-3.44%</td>
</tr>
<tr>
<td>MRI charges ($)</td>
<td>38.94 ± 0.60</td>
<td>57.43 ± 0.62</td>
<td>1.5083662</td>
<td>25.6%</td>
</tr>
<tr>
<td>Surgical procedures §</td>
<td>0.030685 ± 0.000299</td>
<td>0.029715 ± 0.000307</td>
<td>0.0009700</td>
<td>3.16%</td>
</tr>
<tr>
<td>Laboratory charges ($)</td>
<td>376.73 ± 3.16</td>
<td>682.06 ± 3.25</td>
<td>84.67</td>
<td>11.04%</td>
</tr>
<tr>
<td>Total charges ($)</td>
<td>7,281.28 ± 27.85</td>
<td>6,382.92 ± 28.66</td>
<td>898.36</td>
<td>12.34%</td>
</tr>
<tr>
<td>Coronary care charges ($)</td>
<td>196.04 ± 2.12</td>
<td>171.73 ± 2.18</td>
<td>24.31</td>
<td>12.40%</td>
</tr>
<tr>
<td>Radiology charges ($)</td>
<td>436.42 ± 1.90</td>
<td>379.79 ± 1.95</td>
<td>56.63</td>
<td>12.98%</td>
</tr>
<tr>
<td>Pharmacy charges ($)</td>
<td>710.15 ± 4.87</td>
<td>615.54 ± 5.01</td>
<td>94.61</td>
<td>13.32%</td>
</tr>
<tr>
<td>ICU, general claims</td>
<td>0.00879 ± 0.0001</td>
<td>0.007477 ± 0.000103</td>
<td>0.0013127</td>
<td>14.93%</td>
</tr>
<tr>
<td>Radiology other imaging claims</td>
<td>0.015039 ± 0.00013</td>
<td>0.012313 ± 0.000134</td>
<td>0.0027263</td>
<td>18.13%</td>
</tr>
<tr>
<td>General drugs and/or IV therapy claims</td>
<td>0.109558 ± 0.000506</td>
<td>0.089017 ± 0.000521</td>
<td>0.0205408</td>
<td>18.75%</td>
</tr>
<tr>
<td>ICU, medical claims</td>
<td>0.001252 ± 0.00035</td>
<td>0.000889 ± 0.00036</td>
<td>0.0003639</td>
<td>29.06%</td>
</tr>
<tr>
<td>ICU, surgical claims</td>
<td>0.000636 ± 0.000031</td>
<td>0.000416 ± 0.000032</td>
<td>0.000201</td>
<td>34.59%</td>
</tr>
<tr>
<td>Informational encounters</td>
<td>0.000054 ± 0.0000035</td>
<td>0.000029 ± 0.00003</td>
<td>0.0000252</td>
<td>46.49%</td>
</tr>
<tr>
<td>Radiology nuclear medicine claims</td>
<td>0.00141 ± 0.000082</td>
<td>0.000581 ± 0.000085</td>
<td>0.0008288</td>
<td>58.79%</td>
</tr>
<tr>
<td>ICU, psychiatric claims</td>
<td>-0.000193 ± 0.000017</td>
<td>-0.000322 ± 0.000017</td>
<td>0.0001288</td>
<td>66.63%</td>
</tr>
<tr>
<td>Ambulatory surgical care claims</td>
<td>-0.000245 ± 0.000022</td>
<td>-0.000412 ± 0.00023</td>
<td>0.0001672</td>
<td>68.15%</td>
</tr>
<tr>
<td>ICU, other claims</td>
<td>0.0000269 ± 0.000022</td>
<td>0.000062 ± 0.00023</td>
<td>0.0002071</td>
<td>76.84%</td>
</tr>
<tr>
<td>Coronary care unit, heart transplant claims</td>
<td>0.000013 ± 0.000002</td>
<td>-0.000004 ± 0.000002</td>
<td>0.0000165</td>
<td>127.82%</td>
</tr>
<tr>
<td>ICU, pediatric claims</td>
<td>0.000018 ± 0.000002</td>
<td>-0.000013 ± 0.000002</td>
<td>0.0000035</td>
<td>174.51%</td>
</tr>
<tr>
<td>Coronary care unit, other claims ‡</td>
<td>-0.00002 ± 0.000018</td>
<td>-0.000059 ± 0.000018</td>
<td>0.0000387</td>
<td>189.87%</td>
</tr>
<tr>
<td>Coronary care unit, myocardial claims</td>
<td>-0.000007 ± 0.000006</td>
<td>-0.000031 ± 0.000006</td>
<td>0.0000242</td>
<td>351.30%</td>
</tr>
</tbody>
</table>

†: P-value of 0.029. ‡: P-value of 0.096. §: P-value of 0.024. □: P-value of 0.126.

Comparison of components of care in all care settings as available in the Master Beneficiary Summary File (MBSF)

The Master Beneficiary Summary File (MBSF) has an important advantage over all other files as it provides data already aggregated, per beneficiary, from the entirety of the CMS claims regardless of whether your specific Data Use Agreement with CMS covers the files from which the claims were retrieved. For example, our DUA does not cover Durable Medical Equipment claims, notwithstanding their aggregated events and costs are available, albeit at year-level granularity per beneficiary, in the MBSF. This also means that the MBSF provides the most reliable measure of total costs, which, as depicted in Table 4, we found to have increased faster in the ACO cohort than in the non-ACO cohort ($390.77 greater increase in average per beneficiary). In some cases, the non-ACO cohort saved money in comparison to its own baseline, while the ACO cohort did not.

Table 4. Components of care and total costs across all care settings. Numbers are differences in average per beneficiary between the baseline and outcome periods.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Non-ACO</th>
<th>ACO</th>
<th>Difference</th>
<th>Percent diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambulatory Surgery Center Medicare Payments ($)</td>
<td>7.2 ± 0.27</td>
<td>36.17 ± 0.28</td>
<td>-28.97</td>
<td>-802.08%</td>
</tr>
<tr>
<td>Durable Medical Equipment Medicare Payments ($)</td>
<td>-8.14 ± 0.46</td>
<td>6.45 ± 0.47</td>
<td>-14.6</td>
<td>-179.25%</td>
</tr>
<tr>
<td>Other Procedures Medicare Payments ($)</td>
<td>57.98 ± 0.83</td>
<td>151.64 ± 0.85</td>
<td>-93.66</td>
<td>-161.55%</td>
</tr>
<tr>
<td>Part B Physician Events</td>
<td>1.038 ± 0.004</td>
<td>2.643 ± 0.004</td>
<td>-1.605</td>
<td>-154.54%</td>
</tr>
<tr>
<td>Tests Medicare Payments ($)</td>
<td>33.5 ± 0.27</td>
<td>78.61 ± 0.28</td>
<td>-45.11</td>
<td>-134.66%</td>
</tr>
<tr>
<td>Dialysis Events</td>
<td>0.03 ± 0.001</td>
<td>0.067 ± 0.001</td>
<td>-0.037</td>
<td>-123.65%</td>
</tr>
<tr>
<td>Ambulatory Surgery Center Events</td>
<td>-0.061 ± 0.001</td>
<td>0.014 ± 0.001</td>
<td>-0.075</td>
<td>-122.18%</td>
</tr>
<tr>
<td>Imaging Medicare Payments ($)</td>
<td>-28.84 ± 0.2</td>
<td>2.99 ± 0.21</td>
<td>-31.83</td>
<td>-110.35%</td>
</tr>
<tr>
<td>Other Procedures Events</td>
<td>0.795 ± 0.009</td>
<td>1.619 ± 0.009</td>
<td>-0.824</td>
<td>-105.73%</td>
</tr>
<tr>
<td>Tests Events</td>
<td>2.486 ± 0.01</td>
<td>4.994 ± 0.01</td>
<td>-2.509</td>
<td>-100.93%</td>
</tr>
<tr>
<td>Part B Drug Events</td>
<td>0.48 ± 0.004</td>
<td>0.914 ± 0.004</td>
<td>-0.434</td>
<td>-90.52%</td>
</tr>
<tr>
<td>Part B Physician Medicare Payments ($)</td>
<td>113.21 ± 0.23</td>
<td>207.38 ± 0.23</td>
<td>-94.17</td>
<td>-83.18%</td>
</tr>
<tr>
<td>Imaging Events</td>
<td>0.517 ± 0.003</td>
<td>0.928 ± 0.003</td>
<td>-0.411</td>
<td>-79.49%</td>
</tr>
<tr>
<td>Anesthesia Medicare Payments ($)</td>
<td>11.06 ± 0.09</td>
<td>19.46 ± 0.09</td>
<td>-8.39</td>
<td>-75.87%</td>
</tr>
</tbody>
</table>
Anesthesia Events 0.077 ± 0.001 0.134 ± 0.001 -0.057 -73.37%
Durable Medical Equipment Claims 0.367 ± 0.003 0.614 ± 0.003 -0.247 -67.28%
Part B Drug Medicare Payments ($) 213.48 ± 2.11 277.51 ± 2.17 -64.02 -29.99%
Part D Medicare Payments ($) 871.95 ± 3.56 1051.71 ± 3.66 -179.76 -20.62%
Part D Total Prescription Costs ($) 1248.55 ± 3.82 1486.8 ± 3.93 -238.25 -19.08%
Part D Fill Count † 20.732 ± 0.017 23.903 ± 0.017 -3.171 -13.29%
Total Beneficiary Payments ($) 805.88 ± 1.58 889.37 ± 1.62 -83.49 -10.36%
Hospital Outpatient Medicare Payments ($) 815.16 ± 2.7 897.53 ± 2.77 -82.37 -10.10%
Total Medicare Payments ($) 5,950.14 ± 12.35 6,340.91 ± 12.71 -390.77 -6.57%
Evaluation & Management Medicare Payments ($) 223.81 ± 0.56 238.32 ± 0.58 -14.52 -6.49%
Dialysis Medicare Payments ($) 9.1 ± 0.09 8.43 ± 0.09 0.67 7.32%
Other Part B Carrier Medicare Payments ($) 101.44 ± 6.64 93.83 ± 6.05 7.61 7.50%
Emergency Room Visits (inpatient or outpatient) 0.237 ± 0.001 0.217 ± 0.001 0.021 8.74%
Hospital Outpatient Visits 2.664 ± 0.009 2.419 ± 0.009 0.245 9.18%
Home Health Medicare Payments ($) 304.466 ± 1.107 261.3 ± 1.14 39.17 13.04%
Home Health Visits 1.874 ± 0.009 1.529 ± 0.009 0.345 18.42%
Hospice Medicare Payments ($) 240.12 ± 1.21 153.48 ± 1.24 86.64 36.08%
Skilled Nursing Facility Covered Days 1.408 ± 0.005 0.89 ± 0.005 0.518 36.80%
30-day Hospital Readmissions 0.016 ± 0.0002 0.01 ± 0.0002 0.006 36.99%
Hospice Covered Days 1.58 ± 0.008 0.986 ± 0.008 0.595 37.63%
Skilled Nursing Facility Medicare Payments ($) 577.46 ± 2.23 357.97 ± 2.29 219.49 60.01%
Inpatient Covered Days 0.567 ± 0.003 0.347 ± 0.003 0.22 38.84%
Other Inpatient Medicare Payments ($) 196.04 ± 2.1 197.65 ± 2.16 1.61 0.33%

† Defined by the CMS as the number of 30 days of individual drug fills. 12 months of one drug produce fill count of 12. 12 months of 2 drugs produce 24. For more information, see the ResDAC website11.

Discussion

For the most part our data suggests that the cost for the care given to beneficiaries assigned to MSSP ACOs increased faster from the baseline period (2010 plus 2011) to the outcomes period (2013 plus 2014) than the cost of a control cohort, even after adjusting for a total of 69 covariates in a doubly robust estimator – propensity score weighting followed by re-adjustment in a generalized linear regression model. The covariates included spanned beneficiary personal traits, medical history, ACO administrative region, and socioeconomic factors of the beneficiary’s ZIP code of residence, in correspondence to the covariates used in previous studies17,22.

It is expected to see generalized increases in average number of claims per patient, since the fixed cohorts will have aged from 2010 to 2014; and it is expected to find generalized increase in costs since, on top of the increased number of claims, the monetary values are not adjusted for inflation. It is hard to explain, though, why the ACO cohort exhibited net greater increase in total cost than the non-ACO cohort.

Our results are different than those from studies of the Pioneer ACO program17,18 and the Blue Cross Blue Shield of Massachusetts’ Alternative Quality Contract1,2,4,6,7, and the recent analysis of the first year (2013) of the Medicare Shared Savings Program22. Our results suggest an opposite effect which we cannot easily explain. McWilliams et al.12 identified savings among the MSSP ACOs with agreement start date earlier than January 1st, 2013, but not among the ones that started precisely on January 1st, 2013, suggesting that the capacity to generate savings might take some time to be developed. In our study, however, after ignoring year 2012 and analyzing the claims of all 333 MSSP ACOs up until the end of 2014, we found negative savings. Moreover, it may also be worthy of notice that McWilliams et al. emulated as possible the CMS’ algorithm for retrospectively assigning beneficiaries to MSSP ACOs using the official list of MSSP ACO providers, and did so using a 20% random sample. In our study we did not consider the list of MSSP ACO providers. We used only the official list of beneficiaries assigned to MSSP ACOs as provided by CMS via the CMS Virtual Research Data Center (VRDC), and had access to the totality of beneficiaries, albeit our study design drastically reduced the number of beneficiaries considered due to our eligibility criteria.

In an effort to assure complete data for all of our outcomes of interest, we excluded beneficiaries with anything less than complete data in all years under study, as described in the subsection Beneficiary eligibility for this study. The eligibility criteria resulted in the loss of 46.5% of the beneficiaries enrolled in Medicare in all four years considered, and inclusion of only 36.5% of the beneficiaries enrolled in Medicare during the study period that were ever assigned to a MSSP ACO, as depicted in Figure 1. In future analyses we plan to look for unbalances in the patients who were excluded from this study that might help explain our present results.
The figures presented by The Centers for Medicare and Medicaid Services for MSSP SCOs were positive\textsuperscript{30,31} in comparison to the benchmark defined for each MSSP ACO, with $705 million saved by 58 MSSP ACOs in Performance Year (PY) 2013 and $806 million saved in PY 2014 by 92 MSSP ACOs. However, the methods used by the MSSP for quantifying savings are quite different from those of this paper. The MSSP compares actual post-ACO expenditures against ACO-specific projected expenditures\textsuperscript{20} derived from the historical, fee-for-service Medicare Parts A and B expenditures, specific to the beneficiaries assigned to each ACO, while adjusting for beneficiary characteristics and other factors and correcting for the projected growth in national per capita health expenditures. The details of these projection methods can be found in the references\textsuperscript{30}.

Conclusion

Our results suggest that over its first two years the MSSP ACOs may have increased the costs of Medicare spending when comparing to an adjusted control cohort of beneficiaries never assigned to a MSSP ACO, but these results contradict other studies on other types of value-based healthcare reimbursement models, and also a recent study of the first year of the MSSP. Further analysis is needed to understand the differences between our results and previous studies.

Disclaimer

The views and opinions of the authors expressed herein do not necessarily state or reflect those of the U.S. National Library of Medicine, National Institutes of Health or the U.S. Department of Health and Human Services. This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNBC). This research was also supported in part by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DoE) and the NLM.

References


A Connectivity Framework for Social Information Systems Design in Healthcare

Craig E. Kuziemsky, PhD1, Pavel Andreev, PhD1, Morad Benyoucef, PhD1, Tracey O’Sullivan, PhD2, Syam Jamaly, BComm1

1 Telfer School of Management, University of Ottawa, Ottawa, ON, Canada
2 Interdisciplinary School of Health Sciences, University of Ottawa, Ottawa, ON, Canada

Abstract
Social information systems (SISs) will play a key role in healthcare systems’ transformation into collaborative patient-centered systems that support care delivery across the entire continuum of care. SISs enable the development of collaborative networks and facilitate relationships to integrate people and processes across time and space. However, we believe that a “connectivity” issue, which refers to the scope and extent of system requirements for a SIS, is a significant challenge of SIS design. This paper’s contribution is the development of the Social Information System Connectivity Framework for supporting SIS design in healthcare. The framework has three parts. First, it defines the structure of a SIS as a set of social triads. Second, it identifies six dimensions that represent the behaviour of a SIS. Third, it proposes the Social Information System Connectivity Factor as our approximation of the extent of connectivity and degree of complexity in a SIS.

Introduction
Healthcare systems worldwide are undergoing innovative transformation into networked systems that deliver collaborative patient centered care delivery across the continuum of care1. The design of healthcare systems has not been traditionally focused on facilitation of integrated and collaborative care delivery and, therefore, health information systems (HISs) will play a key role in transformative efforts2, 3. Social information systems (SISs), defined as “information systems based on social technologies and open collaboration”4, will play a particularly significant role in these transformative efforts because of their ability to facilitate and support social processes of healthcare delivery5, 6. SISs have a different focus than traditional HISs since their purpose is to facilitate collaboration and networking over time as opposed to the automation of specific tasks such as order entry7. The fundamental tenet of SISs are the need to support and nurture the development of social and collaborative competencies over time. Therefore, the scope of SISs is much broader than task based HISs such as decision support or order entry systems as SISs must facilitate provision of a service (e.g. patient participation, community resilience) while also developing the social competencies inherent within these systems. Although we adopt the definition of SIS proposed by4 we believe that SISs in healthcare present more challenges compared to SIS design in general. These challenges are related to the extent of openness, transparency, and open collaboration, the hallmark of SISs4 which are more complex in healthcare compared to other sectors.

Despite SISs being a recent phenomenon they have shown potential for redefining the way in which clinical care or public healthcare can be provided. Examples of SISs can be seen across the healthcare spectrum. Coiera8 has described how social relationships and networks can lead to new approaches for diagnosing and managing illness. At the micro (clinical) level, social networking tools like Dr. Google and PatientsLikeMe® provide information to patients about illness and treatments and enable connectivity for patients’ discussion groups. At a macro (public health) level, activities such as disaster management and community resilience have benefited from social networking applications because they have enabled social interaction and collaboration that is a key aspect of public health activities9, 10. However, despite their benefits, there are also challenges in designing and implementing SISs that must be properly addressed. At a micro level, simply providing patients with access to data in inappropriate contexts may overwhelm patients and provide overload from the information and opinions that they receive through support groups11. Similarly, at a macro level, it has been challenging to develop SISs due to the disparate and multidisciplinary nature of the communities where these systems are used, and the range of user needs within SISs12.
One essential lesson that we have learned from the implementation of HISs is that unintended consequences will arise post implementation. These unintended consequences occur for a number of reasons but a common issue is changes to work practices resulting from technology mediated connectivity, people, and processes. SISs and the connectivity they enable provide a bigger base for unintended consequences to occur due to the web of connections enabled by SISs. Before we can manage unintended consequences we first need to understand how they originate. However, this is particularly challenging in SISs because of what we refer to as a “connectivity” issue, which refers to the scope and extent of system requirements for a SIS. SISs create connectivity that is unparalleled, and further coupled with the fact that there are different users and patterns of use, SIS connectivity can be a substantial design problem.

We define “connectivity” as the organizations, tasks, people and technology needed to achieve an outcome. Connectivity is a common challenge in HIS design and thus there have been several attempts to understand it in HIS design. Sociotechnical approaches for systems design such as Sittig and Singh’s eight step Sociotechnical Model defines sociotechnical “connectivity” by such dimensions as people, workflows, technologies, and policies associated with designing, implementing, and using HIS. Participatory design facilitates “connectivity” in design by engaging end users and binding design in user needs. Behavioral design approaches like Task Technology Fit (TTF) or Cognitive Task Analysis (CTA) use tasks (e.g., decision making needs) to bind system design parameters. Outside of healthcare, Supply Chain Management (SCM) defines system requirements according to business models that articulate the necessary connectivity to facilitate supply chain activities. For example, SCM has a common reference model called the Supply Chain Operations Reference (SCOR) model with five general processes (plan, source, make, deliver and return). The SCOR model has been used to understand supply chain connectivity for IS design and the development of metrics.

While existing SIS research has developed frameworks for technical architectures and identified challenges to developing SISs, studies on how to model connectivity as a means of SIS design and evaluation do not exist. SIS design in healthcare is particularly challenging as it requires the connection of people, processes and technology across different workflows and user roles. Healthcare SIS design requires new approaches that enable us to understand complex connectivity as a precursor to systems design. Coiera proposed Interaction Design Theory (IDT) as a systems design approach that focuses less on individual issues or technologies and instead tries to understand the web of interactions that exist within a system. IDT allows designers to make predictions about how a group as a whole will interact in complex settings.

However, there is limited work that has looked at SIS design in healthcare to identify specific elements of social connectivity and how they impact systems design. The main contribution of this paper is the Social Information System Connectivity Framework for SIS design in healthcare. The framework has three parts. First, it defines the structure of a SIS as a set of social triads. Second, it identifies six dimensions that represent the behavior of a SIS. Third, it proposes the Social Information System Connectivity Factor as our approximation of the extent of connectivity and degree of complexity in a SIS. The paper has four sections. Section 1 was the introduction. Section 2 is the materials and methods. Section 3 is the description of the Social Information System Connectivity Framework for supporting SIS design in healthcare. We conclude with a discussion of the implications of our work and the next steps from the research presented in this paper.

Materials and Methods

Data Sources

Two data sources informed our study. First, the authors have studied HIS design to support social and collaborative health delivery at clinical and population health levels including studies on perioperative systems, palliative care systems, and community resilience as part of designing disaster management systems. The common thread across all these studies was to understand different types of collaboration, and how it informs HIS design to support the development and maintenance of collaborative and social practices. The case studies involved user engagement methods such as community based participatory research and participatory design. In the context of these methods we spent considerable time with users in order to gain an appreciation of social needs and competencies and how they develop over time.

The second data source was a literature search on SISs, networks, and modelling approaches for SIS design in general and in healthcare. Scopus, IEEE Xplore, and the ACM Digital Library were searched. The literature search provided additional insight on SISs, in particular frameworks for representing and modeling them.
Data Analysis

We used descriptive qualitative content analysis on the literature we retrieved and the integrated findings from our case studies. Our objective was to integrate the empirical data from our studies with the literature on SIS modeling and design to develop a general framework on SIS design in healthcare. To provide a framing for our analysis we drew upon a paper that described how accountable healthcare delivery must be viewed from the perspective of a structure and a set of behaviours. We adopted that framework as we analyzed our data so our analysis identified structural and behavioral aspects of SISs.

Results

Our results are presented in two sections. First, we describe the connectivity framework for SIS design. Second, we present the three specific components of the framework corresponding to the structure and behavior, and third, we describe operationalization of the framework as the Social Information System Connectivity Factor.

Connectivity Framework for SIS Design

Figure 1 shows our connectivity framework for SIS design in healthcare. The framework addresses the previously described shortcomings in the literature such as the need to represent healthcare delivery as a complex ecosystem that can include multiple actors, settings and information flows that may be subject to different degrees of governance. A first step towards understanding connectivity based SIS design is to decompose the components of the system so they can be modelled. Our framework is labeled with three parts: structure, behavior and operationalization as the Social Information System Connectivity Factor (SISCF). Each part is described in detail below the figure.

![Connectivity Framework for SIS Design in Healthcare](image)
Connectivity Framework – Structure

The structure of the connectivity framework represents a social ecosystem as a number of interrelated social triads with three concepts: person, process and technology. Those three concepts were common in the literature we retrieved on SISs in healthcare as well as in social modelling approaches. Studies on social networks or technologies to support healthcare delivery most often focused on people, processes (e.g., information exchange, communication, care coordination), and technology that is used.

From the literature review on social modeling we identified social modelling languages such as Social Business Process Model and Notation (SBPMN) that emphasize modelling of social activities such as community generated events and social relationship links. The social triad aspect of our framework represents the necessary connectivity for SIS design. A social triad can be at the micro level of an individual patient where the extent of social linkages would be at the level of processes and interactions that a patient uses as part of his/her social network for managing his/her own chronic illness. In that example, the connections we would be interested in modelling would be the clinicians (e.g., physician, pharmacist, dietician) that are part of the patient’s social ecosystem as well as the processes (e.g., glucose or diet monitoring, exercise regime) and communication that a patient goes through as part of his/her disease management. If we move to meso level care delivery, we would be linking together two or more social triads such as for the integration of multiple units in a hospital, or multiple clinics that work together as part of providing integrated collaborative community care delivery. The connections we would be interested in modelling at that level would include collaborative care processes (e.g., joint decision making) and integrated data sharing. At the macro level, we would be integrating multiple social triads as part of modelling population health initiatives such as community resilience to support disaster management, or public health initiatives such as managing an influenza outbreak. The connections we would be interested in modelling at the macro level would include collaboration, interoperability and governance across communities and organizations.

Connectivity Framework - Behavior

The behavioral component of our framework defines how the social triads (e.g., structure) work for a particular SIS. From our analysis of the two data sources we identified six behavioral dimensions relevant to SIS design in healthcare: user driven design flexibility, empowerment and responsibility, workflow extensibility, process immaturity, data standards & interoperability, and governance complexity. The behavioral dimensions span a range of considerations from workflow and motivation to use a SIS, to system interoperability and governance issues. Below we discuss the six behavioral dimensions and how they help us understand the degree of complexity as part of SIS modelling and design.

User Driven Design Flexibility

One of the key challenges of SISs is that they are driven by the needs or requirements of users. This introduces two sets of challenges. At a micro level, patient participatory medicine is dependent on the willingness of people to be the active stewards of their own care delivery. Further, SISs to support patient participation can range from basic information retrieval to information exchange and partnership and collaboration on the content. Those different tasks require very different design solutions. SISs require much more flexibility due to the nature of how people conduct tasks. Similarly, at macro levels it has been shown that the diversity of end users necessitates consideration of the needs of diverse individuals. While the premise of SISs is connectivity and socialization across human networks, achieving that goal makes user driven design more challenging. To address that issue, the diverse users of SIS and their needs must be identified and used to inform systems design. However, a social community may have users with a range of technological skills and that needs to inform systems design. One of our case studies on SIS design to support community resilience for disaster management highlighted that system design requirements have to start technologically at the lowest common denominator to enable all users to become engaged and comfortable with using the system.

Empowerment and Responsibility

Unlike HISs designed to support a specific task, SISs are dynamic systems that are intended to develop and maintain relationships over time. That puts the burden of responsibility on system users to maintain activity in the SIS over time to enable it to grow and develop. SISs provide a means of empowering people, but empowerment brings with it certain responsibilities. At a micro level, patients recovering from an illness see something explicit in using a SIS to guide their therapy and disease recovery. However, people may be less motivated to use it for data collection for routine monitoring (e.g., blood pressure measurement, diet). Research has shown that a relatively low percentage of patients adequately document the necessary data for disease management and also that while patients may initially be
very keen to collect illness data, the enthusiasm wanes over time\(^9\). Macro level public health initiatives have a similar challenge in that community resilience efforts to support disaster management relies on a mixture of private and public and paid and volunteer workers to maintain progress\(^{23}\). A key public health issue is that disaster management or disease surveillance are preparing for events which may never happen and thus keeping parties motivated to continue to use and maintain a SIS can be a challenge\(^{12, 33}\).

**Workflow Extensibility**

A challenge with socially driven care delivery is the processes that are done may lack rules of engagement for how they are to be conducted. In our case studies we found that workflows for social processes were often poorly defined at both micro and macro levels. At a micro level, patient participatory medicine requires the creation of new workflows to accommodate both patients and clinicians with respect to information exchange and decision making\(^{29, 34}\). As more care delivery moves outside traditional settings such as hospitals and into the community (e.g. initiatives such as aging in place) it adds more people such as informal caregivers to a social triad which adds further workflow complexity. Therefore, workflow modelling must be done in a way that takes into perspective all of these diverse user groups\(^{39}\). Similarly, workflows are often undefined for community level initiatives such as disaster management that can necessitate the need for workflows that span micro (individual) and macro (community) perspectives\(^{36, 37}\).

**Process Immaturity**

While relationships, social connectivity, and collaboration are the tenets of SISs, a challenge is that many of the social processes that we are trying to implement are still maturing and may be in an evolutionary state. At the micro level, patient centred care and patient participatory medicine are evolving concepts\(^{34}\). Similarly, collaborative team based care delivery at the meso level has been described as existing more in name than in actual implementation\(^{38}\). The rules of engagement for how collaboration needs to occur have to be better understood and defined before we can expect a SIS to implement and foster collaborative care delivery\(^{39, 40}\). One of our macro level case studies described how the common ground needed for social collaboration in a community goes through a development cycle where people need to first develop coordination and communication practices before they can collaborate\(^{24}\).

**Data Standards & Interoperability**

Interoperability across different settings is an essential requirement of HISs\(^{41}\). A key driver of HIS interoperability has been the development of standards for data exchange (e.g., HL7) and terminology (e.g., SNOMED). On one hand, SISs offer the potential for social triad interoperability and data collection beyond traditional healthcare settings and systems, but they also introduce connectivity complexity to traditional interoperability standards. As we use more social media applications (e.g., Facebook) or self-monitoring tools (e.g., Fitbits, Smartphones) as sources of healthcare data, it threatens to erode the extensive work that has been done in developing formal interoperable healthcare data standards. In recognition of this issue, recent research has suggested the need to change the focus from developing formal data standards to developing tools to enable extraction and analysis of social media application and self-monitoring application data\(^{42, 43}\).

**Governance Complexity**

Governance complexity refers to the need to consider activities within the larger social structure where they occur, including the relationship across entities (e.g., organizations) that may impact social structures by exerting influence or autonomy\(^{24}\). SISs cross different units both within (i.e., intra-organizational) and across (i.e., inter-organizational) settings. The more social triads that are integrated as part of healthcare delivery, the greater the governance complexity challenges in the form of cross-organizational information sharing and the need to integrate different types of agents, policies, and procedures\(^{23, 45}\). Governance issues can at times be a significant impediment to the development of social relationships, at both micro and macro levels, an example being the inability to share necessary information to support social healthcare processes\(^{24, 44}\).

**Connectivity Framework - Operationalization**

To operationalize the connectivity framework, we introduce the Social Information System Connectivity Factor (SISCF) – bottom of fig.1 - as our approximation of the extent of connectivity in a SIS. The SISCF is comprised of two elements, the connectivity complexity (CC), and the connectivity time (CT). The CC helps us approximate and understand the complexity of the structural and behavioral components within a SIS. For example, connectivity is more complex for a meso level SIS that integrates five social triads compared to a SIS integrating two triads because there will be more connections and relationships to consider. As more behavioral dimensions (e.g., governance across multiple settings, different workflows, disparate forms of social data, needs of a variety of end users) are incorporated
into a SIS, the degree of complexity also increases. It must also be emphasized that there is no one pattern of alignment for the structural and behavioral dimensions in our framework. Sittig and Singh point out that sociotechnical frameworks in healthcare need to be viewed as complex adaptive systems and that framework dimensions must be studied in a non-linear manner with an emphasis on how different dimensions interact and relate to each other. An implication of implementing multi-dimensional connectivity is that trade-offs will have to occur in many of the behavioral dimensions. In one of our case studies we showed that providers and administrators can have different needs with respect to how a collaborative HIS is used. Administrators wanted data to support organizational decision making on workloads and to provide necessary reporting for healthcare accreditation bodies, but collecting that data increased the data entry workload of the front line clinicians. In that situation the workflow and governance behavioral dimensions were in conflict with each other. Similarly, establishing the level of common ground necessary for social processes often means that individuals may have to change how they do things in consideration of social good. Again, that requires behavioral changes at the individual level as a precursor to achieving social connectivity at a group level.

The CT refers to the length of time that social connectivity has taken place, as some of the above CC dimensions are influenced by temporal properties. A key time related factor is the maturity of processes. Many of the social processes that are done in healthcare such as collaborative healthcare delivery or patient participatory medicine are still maturing and thus will develop over time. Even if a SIS is designed to support social processes such as collaboration and coordination there may be a stepwise progression to using it to its full capacity. One of our case studies showed that even though we designed a palliative care system with alerting and reminder features to support collaborative care delivery they were not used to capacity because the technology was more advanced than the care processes that were using it. A different study on HIS implementation described how users initially used an EHR system solely as a documentation tool but over time they developed and adapted social coordination strategies that were enabled through the EHR. Social technologies are multi-purpose tools that can provide a range of social functionality depending on the degree of social processes which are using them. For example, Twitter and Facebook were initially designed for social communication at a personal level but they evolved in use to include clinical and public health activities. The first step in the development of SISs must be an analysis of the behavioral dimensions of a social system using the above described behavioral dimensions as there is little point in implementing SISs that are more advanced than the people and processes using them. Once the social process baseline has been identified, an SIS can then be designed, which must include tools and features to support social processes in the present while also helping the processes mature and evolve over time.

**Discussion**

SISs offer great potential for improving healthcare delivery in that they can support the development and ongoing maintenance of relationships to support collaborative healthcare activities. Managing chronic disease at the micro level is dependent on our ability to connect a patient with information, processes, people and technology as part of disease management. Similarity, public health efforts like disaster management and community resilience are dependent on relationship building over time that encompasses a wide variety of user groups. However, SISs introduce a connectivity problem in that the system design considerations, and potential UICs of systems design, are so vast that SIS design becomes a significant challenge. This research helped to address the above issue by proposing the Social Information System Connectivity Framework for SIS design. The framework defines the structure of a SIS as a set of one or more social triads and the behaviour of a SIS according to six behavioural dimensions. The framework is operationalized through the Social Information System Connectivity Factor and the connectivity complexity (CC) and connectivity time (CT) concepts, which help us understand how the degree of connectivity and temporal aspects of social connectivity will impact SIS modelling and design.

SISs present a new perspective on the design and evaluation of health information systems. We do not believe it is possible to eliminate all unintended consequences from SIS implementation due to the complexity of the healthcare domain. Rather our focus should be on identifying and managing unintended consequences. The value of our framework is that it enables us to proactively identify and understand the connectivity complexity of SIS design. We emphasize that we do not provide a prediction of SIS connectivity complexity but rather provide a way of understanding the complexity and temporal considerations associated with it. This research is complementary to existing work that has developed architectures for SIS design or identified challenges in performing SIS design as well as to system design approaches like Interaction Design Theory. While technical aspects of SISs are essential for designing system infrastructures, perhaps a bigger challenge in SIS design are the behavioural dimensions identified in this paper. For example, processes like patient self-disease monitoring (e.g., blood pressure, diet) and community preparation for disaster management are both preparing for events that may not occur for some time (i.e.,
illness, natural disaster), if at all. Therefore, keeping users motivated to keep a SIS active in the present moment, as well as enabling it to evolve and mature, presents a significant challenge. Similarly, while the value of SISs comes from the social network of users within it, it also introduces connectivity challenges with respect to adaptiveness of workflows to address the dynamic nature of the system, incorporating social data standards, and managing cross organizational interoperability and governance. A significant implication of multi-dimensional connectivity is that SIS design cannot satisfy everyone, and therefore trade-offs will have to be made with respect to the behavioral dimensions of a social system. An example of such a trade-off is in the governance complexity dimension where traditional information sharing agreements that impede cross organizational information sharing will need to be loosened to enable development of social networks. Trade-offs at the individual level as a precursor to achieving common ground at a group level is another example of a trade-off as part of social connectivity.

One significant finding from our work is that the development trajectory of SISs is understated. While social technologies are well developed from a tool perspective, the social processes that use the tools are not nearly as well developed. Therefore, our effort moving forward needs to focus on understanding how to design SIS to provide the necessary connectivity to support present needs but also to enable the ongoing development of social processes. Bottom line is we cannot implement SISs without ensuring that the social processes that will use them have sufficient maturity to benefit from the implementation. Another implication for SIS design from our framework is the need to study how different social networking tools contribute to a SIS. Self-monitoring tools such as Fitbits are fairly bounded in terms of the data they collect (e.g. steps, heartrate) compared to tools such as Twitter or Facebook that are far less bounded in terms of data collected. Thus, there is a need to study connectivity implications for these different data sources. Further, increased sharing of data brings about privacy and security issues, and, while not discussed explicitly in this paper, they are important considerations in SIS design.

Limitations of our research are that we have developed but not used the connectivity framework to model or design SISs. The next stage of our research is to use the framework to design and evaluate SISs at the clinical (micro and meso) and community health (macro) levels.

**Conclusion**

SISs will play a crucial role in healthcare transformation at all levels (micro, meso and macro). Therefore, there is a need for approaches that help us to understand and model the complex connectivity that is inherent in a SIS. This paper proposed the Social Information System Connectivity Framework for SIS design. This research helped to address the above issue by proposing the Social Information System Connectivity Framework for SIS design. The framework defines the structure of a SIS as a set of one or more social triads and the behavior of a SIS according to six behavioral dimensions. The framework is operationalized through the Social Information System Connectivity Factor and the connectivity complexity (CC) and connectivity time (CT) concepts, which help us understand how the degree of connectivity and temporal aspects of social connectivity will impact SIS modelling and design.

**Acknowledgements**

We acknowledge funding support from the Natural Sciences and Engineering Research Council of Canada.

**References**

Abstract

The Zika virus (ZIKV) outbreak in South American countries and its potential association with microcephaly in newborns and Guillain-Barré Syndrome led the World Health Organization to declare a Public Health Emergency of International Concern. To understand the ZIKV disease dynamics and evaluate the effectiveness of different containment strategies, we propose a compartmental model with a vector-host structure for ZIKV. The model utilizes logistic growth in human population and dynamic growth in vector population. Using this model, we derive the basic reproduction number to gain insight on containment strategies. We contrast the impact and influence of different parameters on the virus trend and outbreak spread. We also evaluate different containment strategies and their combination effects to achieve early containment by minimizing total infections. This result can help decision makers select and invest in the strategies most effective to combat the infection spread. The decision-support tool demonstrates the importance of “digital disease surveillance” in response to waves of epidemics including ZIKV, Dengue, Ebola and cholera.

*Corresponding author: eva.lee@gatech.edu

Introduction

Zika virus (ZIKV) is an arbovirus belonging to the family Flaviviridae, and is closely related to the dengue, yellow fever, Japanese encephalitis, and West Nile viruses. It was first isolated from a monkey in the Zika forest of Uganda in 1947. In 1948, the second isolation was made from Aedes africanus mosquitoes in the same forest. Although earlier studies have suggested the possibility of human infection, before 2007, ZIKV rarely caused recognized ‘spillover’ infections in humans, even in highly enzootic areas. Since 1951, human ZIKV infections have been reported in African countries including Uganda, Nigeria, Senegal, and Gabon, and parts of South Asia including India, Malaysia, and Indonesia. In 2007, the first documented ZIKV outbreak was reported from Yap State, Federated States of Micronesia. French Polynesia also reported an outbreak with 28,000 estimated cases in 2013. Since April 2015, Brazil has been experiencing a ZIKV outbreak. The outbreak has subsequently spread to other countries in South America, Central America, and the Caribbean.

ZIKV is primarily transmitted by the bites of infectious Aedes mosquitoes, especially Aedes aegypti and Aedes albopictus. Non-vector borne transmissions, though rare, have also been reported. Symptoms of ZIKV infection include fever, rash, conjunctivitis, and retro-orbital pain. These symptoms are usually mild and unspecific. The ZIKV infection is difficult to detect since its symptoms are similar to those of dengue and chikungunya. Asymptotic cases are also frequent, which makes it even harder to precisely diagnose the ZIKV infection. Although the symptoms of ZIKV infections are usually mild and unpleasurable, it has recently been connected to microcephaly in infants born to mothers infected with ZIKV during pregnancy, suggesting the possibility of intrauterine infection. ZIKV is also suspected to be linked to Guillain-Barré Syndrome (GBS), a muscle weakness caused by the immune system damaging the peripheral nervous system.

With the ease of modern transportation, there is a potential that the ZIKV outbreak may become a global public health crisis. Therefore, it is crucial to understand the transmission dynamics of ZIKV and establish a set of strategies that help to contain the outbreak effectively. Mathematical models have long been used to analyze infectious disease outbreaks to facilitate development of policies and strategies for understanding disease propagation and containment strategies. Since ZIKV is closely related to West Nile, dengue, and yellow fever viruses, it can be modeled with similar techniques. Because of dengue’s global impact, its disease dynamics have been studied extensively, and compartmental models based on Susceptible, Infectious, Recovered (SIR), and Susceptible, Exposed, Infectious, Recovered (SEIR) hosts have been proposed and analyzed. The population dynamics and life table of the major vector of ZIKV, Aedes mosquitoes, are also well studied, and can be incorporated in the compartments for vectors.

In this work, we present a compartmental model for ZIKV based on SEIR host model and SEI vector model, using logistic growth in human population and dynamic vector population. We conduct sensitivity analysis on multiple parameters to understand the risk and transmission of ZIKV, and evaluate the efficiency of different containment strategies. We discuss how this may assist in decision making to control and mitigate outbreaks.
Method and Design

The compartmental model consists of two parts: the compartments for human, subscripted with $H$; and compartments for vectors, subscripted with $V$. Although there have been a few reported deaths related to ZIKV, most ZIKV cases only result in mild symptoms among infectious individuals. We use an SEIR (Susceptible, Exposed, Infectious, Recovered) model to describe the disease spread among the human population. Let $S_H(t), E_H(t), I_H(t)$ and $R_H(t)$ denote the number of individuals in each compartment at time $t$. The human population at time $t$ is given as $N_H(t) = S_H(t) + I_H(t) + R_H(t)$. During the outbreak, the human population is assumed to follow a logistic growth with carrying capacity $K_H$.

The compartments for vectors consist of both the infectious status and the life stages of the mosquito. In particular, a typical life cycle of a mosquito has four stages: egg, larva, pupa, and adult. Let $V_1(t), V_2(t)$ and $V_3(t)$ denote the number of eggs, larvae, and pupae of vectors in the environment. Since only female adults bite humans and transmit the virus, there are no infectious compartments for these three stages. Let $S_V(t), E_V(t)$ and $I_V(t)$ denote the number of susceptible, exposed, and infectious female adults in the environment at time $t$, and denote the total female vector population as $N_V(t) = S_V(t) + I_V(t) + E_V(t)$. Due to their short life cycle, the birth and death rate of vector population is taken into consideration. The transmission from susceptible to infectious in both human and vector population is associated with mosquito bites. Figure 1 shows the transition diagram.

Let $b$ denote the average biting rate per unit time of a female mosquito: $b$ typically ranges from 0.3 to 0.5 per day\(^{41, 43}\). Let $\beta$ represent the percentage of bites that transmit ZIKV. In addition, let $M$ denote the amount of alternative hosts besides human in the environment. Then the number of bites received by a human per unit time is $bN_V/(N_H + M)$, among which $I_V/N_V$ are from infectious mosquitoes. Therefore, the exposure rate of a human is given by

$$\beta \frac{bN_V}{N_H + M} = \beta b I_V$$

Similarly, the exposure rate of a mosquito is $\beta b I_H/(N_H + M)$. Assume the inverse incubation time of ZIKV for human is $\alpha_H$, and the recovery rate of an infectious individual is $\gamma$. The human population growth rate is $\lambda$, and the decreasing human population due to limited carrying capacity is $\lambda N_H^2/K_H$, and this amount is distributed in the four compartments according to their relative size. Then the transitions between compartments for human can be described by the following system of ordinary differential equations:

$$\frac{d}{dt}S_H = \lambda N_H (1 - \frac{S_H}{K_H}) - \beta b I_V - \frac{\mu_H}{N_H + M} S_H$$
$$\frac{d}{dt}E_H = \beta b I_V + \frac{\mu_H}{N_H + M} S_H - \alpha_H E_H - \frac{\mu_H}{N_H + M} E_H$$
$$\frac{d}{dt}I_H = \alpha_H E_H - \gamma I_H - \lambda N_H \frac{I_H}{K_H}$$
$$\frac{d}{dt}R_H = \gamma I_H - \lambda N_H \frac{R_H}{K_H}$$

For the vector compartments, let $e_V$ denote the number of eggs laid per female mosquito per unit time, and $\alpha_V$ the inverse incubation time of ZIKV in vectors. Let $\tau_i$ and $\mu_i, i \in \{1, 2, 3\}$ be the reciprocal of the development time and mortality rate of vectors in stage $i$, respectively. In addition, let $\mu_V$ denote the mortality rate of adults. Assume $q$ is the percentage of females among adult vectors. Since mosquito larvae live in water with limited resources, the dynamics of the larvae population is also assumed to follow a logistic growth with carrying capacity $K_V$. The mortality rate of larvae caused by limited carrying capacity is denoted as $\kappa$. Then the transitions between compartments of vectors are described by the following system of ordinary differential equations:

$$\frac{d}{dt}V_1 = e_V N_V - (\tau_1 + \mu_1) V_1$$
$$\frac{d}{dt}V_2 = \tau_1 V_1 - (\tau_2 + \mu_2) V_2 - \kappa \frac{V_2}{K_V}$$
$$\frac{d}{dt}V_3 = \tau_2 V_2 - (\tau_3 + \mu_3) V_3$$
$$\frac{d}{dt}S_V = q \tau_3 V_3 - \frac{\beta b I_H}{N_H + M} S_V - \mu_V S_V$$
$$\frac{d}{dt}E_V = \beta b I_H + \frac{\mu_V}{N_H + M} S_V - (\mu_V + \alpha_V) E_V$$
$$\frac{d}{dt}I_V = \alpha_V E_V - \mu_V I_V$$

Figure 1. Transition diagram of the compartmental model. The dashed lines refer to the transitions associated with bites. The lines for vector birth and death are omitted for simplicity.
To quickly contain the outbreak and reach a disease-free equilibrium, it is desirable to reduce the basic reproduction number $R_0$, which is the average number of secondary infections produced by a single infected individual in a completely susceptible population. To achieve this, we may alter the parameters by introducing interventions to contain the outbreak. In particular, the following two strategies are considered:

1. Reduce biting rate $b$ by avoiding mosquito bites, using insect repellents, wearing long-sleeved clothes and long pants, and using air conditioning and window/door screens to keep mosquitoes outside.

2. Reduce adult female vector population or increase adult female vector mortality rate $\mu_V$ at equilibrium $S^*_V$ by widely applying insecticides in areas with high (infectious) mosquito population, or introducing genetically modified Aedes mosquitoes, which will cause its offspring to die by reducing larval survival rate and adult longevity.

Thus, containment is achieved when $R_0 < 1$. In that case, an infected individual produces less than one new infected individual during its infectious period, thus the outbreak will start to contain. Using the $R_0$ formulation, we may alter the parameters by introducing interventions to contain the outbreak. In particular, the following two strategies are considered:

Table 2 summarizes the parameters used in modeling the Zika disease dynamics for Rio de Janeiro State (Sentinel surveillance sites reports) and for a town with 100,000 people. Between January 2015 to July 2015, a total of 364
suspected cases were reported in Rio de Janeiro State. By November, the outbreak was contained with approximately 440 infection cases.

Table 2. Estimated parameter values used in the simulation.

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>Rio de Janeiro State</th>
<th>A town of 100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial human population</td>
<td>$N_H(0)$</td>
<td>16,231,365</td>
<td>100,000</td>
</tr>
<tr>
<td>Initial human infections</td>
<td>$I_H(0)$</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Initial adult vector population</td>
<td>$N_V(0)$</td>
<td>70,000,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Initial exposed adult vectors</td>
<td>$E_V(0)$</td>
<td>5</td>
<td>1,000</td>
</tr>
<tr>
<td>Initial infectious adult vectors</td>
<td>$I_V(0)$</td>
<td>5</td>
<td>1,000</td>
</tr>
<tr>
<td>Initial number of eggs</td>
<td>$V_1(0)$</td>
<td>180,000,000</td>
<td>500,000</td>
</tr>
<tr>
<td>Initial number of larvae</td>
<td>$V_2(0)$</td>
<td>90,000,000</td>
<td>150,000</td>
</tr>
<tr>
<td>Initial number of pupae</td>
<td>$V_3(0)$</td>
<td>70,000,000</td>
<td>100,000</td>
</tr>
<tr>
<td>Number of infectious bites delivered per vector per day</td>
<td>$\beta_b$</td>
<td>0.071</td>
<td>0.09</td>
</tr>
<tr>
<td>Carrying capacity for human</td>
<td>$K_H$</td>
<td>30,000,000</td>
<td>120,000</td>
</tr>
<tr>
<td>Carrying capacity for larvae</td>
<td>$K_V$</td>
<td>100,000,000</td>
<td>5,000,000</td>
</tr>
<tr>
<td>Larvae mortality rate due to limited carrying capacity</td>
<td>$\kappa$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Number of alternative hosts</td>
<td>$M$</td>
<td>100,000</td>
<td>80,000</td>
</tr>
<tr>
<td>Human population growth rate</td>
<td>$\lambda$</td>
<td>0.00005</td>
<td>0.00005</td>
</tr>
<tr>
<td>Recovery time of human</td>
<td>$\gamma$</td>
<td>7 days</td>
<td>7 days</td>
</tr>
<tr>
<td>Incubation time of human</td>
<td>$\alpha_H$</td>
<td>3 days</td>
<td>3 days</td>
</tr>
<tr>
<td>Incubation time of vector</td>
<td>$\alpha_V$</td>
<td>10 days</td>
<td>10 days</td>
</tr>
<tr>
<td>Basic reproduction number (estimate from parameters)</td>
<td>$R_0$</td>
<td>1.1139</td>
<td>1.109</td>
</tr>
<tr>
<td>Simulation start date</td>
<td></td>
<td>Jan 1, 2015</td>
<td>Day 0 (outbreak began)</td>
</tr>
<tr>
<td>Implementation of strategies</td>
<td></td>
<td>May 1, 2015</td>
<td>Day 100</td>
</tr>
<tr>
<td>Time for containment without intervention (when asymptotic behavior is observed)</td>
<td></td>
<td>Nov. 2015</td>
<td>Day 900</td>
</tr>
</tbody>
</table>

**Rio de Janeiro State, Brazil** Figure 2 contrasts the reported cases versus simulation by varying the recovery time ($\gamma$) and the incubation time of human ($\alpha_H$). The results suggest that Table 2 provides good estimates for the reported cases (green solid curve vs red dotted curve). When the recovery time increases from 6 to 8 days (dotted vs dashed green curves), the cumulative infections increases from 340 to 566 by July 2015 (66.5% increase). Longer recovery time also leads to delayed containment. This confirms that recovery time has significant impact on disease containment. Thus timely treatment and vaccination for combating ZIKV are critical.

Varying the human incubation period from 2 days to 4 days has a relatively minor influence on the overall infections (from 427 to 459, 7.5% increase) and the outbreak trend (contrast the three solid lines).

**Figure 2.** Varying recovery time and incubation time for human and its effect on overall infection and containment.

Figure 3 contrasts containment effectiveness for four strategies: reducing biting rate ($b$), increasing adult vector mortality rate ($\mu_V$), reducing larvae carrying capacity ($K_V$), and reducing adult female vector population ($N_V$). Without these strategies, total infection reaches 443 and will be contained by October 2015. In our analysis, we assume each strategy is implemented on May 1, 2015. Figure 3(a) shows that reducing biting rate by a mere 20% will lead to an early containment by August 2015 (90 days after implementation) with 270 total infections (39% reduction). The containment is achieved rapidly (< 60 days after implementation) with total infection under 200 when biting rate is
reduced by at least 40%. This proves that reducing biting rate is highly effective. This strategy is also relatively easy to implement.

Increasing the mortality rate of adult vectors is also effective. Figure 3b shows that increasing the vector mortality rate by 20% will lead to an August containment with total infection of 282 (36% reduction). Containment will be achieved instantly with no more than 220 infections if the mortality rate is increased by at least 40%. While this strategy can be achieved by introducing genetically modified mosquitoes to the environment, it is difficult to realize and economically inferior when compared to other strategies.

The larva carrying capacity has to be reduced by 80% in order for total infection to reduce to 330 (26% reduction) with containment achieved by October (Figure 3c). Nonetheless, examining and clearing the water ponds and humid areas where the larvae live remains an important strategy for public health protection.

Reducing vector population not only affects the number of infections at containment, it also changes the trend of the outbreak (Figure 3d). Specifically, reducing total adult vector population by 20% would postpone the outbreak by a month and reduce total infection to 334. Containment can be rapid when reduction is targeted at 80% (total infection reduces to 130). Figures 3c and 3d highlight that reducing adult population is more effective, though applying insecticide widely may come with numerous environmental and health issues.

![Figure 3](image)

**Figure 3.** Contrast of total infection and outbreak trends for four containment strategies.

In practice, combination strategies are often most effective. For brevity, Figures 4a and 4b report containment results for combination of two strategies only. In both figures, the vertical axis represents the cumulative number of human infections at containment. When 40% reduction is achieved in both biting rate and larvae carrying
capacity, it results in 180 total infections (Figure 4a). Similarly, the combination of reducing vector population by 20% and increasing vector mortality rate by 20% results in 232 total infections (Figure 4b).

Since it can be difficult to achieve high effectiveness using a single strategy, combination strategies are appealing. Our results (Figures 4a and 4b) also display marginal improvement beyond certain threshold values in the strategies. This helps policy makers to invest their resources optimally for best possible gain.

Figure 4. Cumulative number of human infections at containment under combination strategies.

**A Town with 100,000 People** To better understand the behavior and sensitivity of our model and the disease spread characteristics, we perform the analysis on a town with 100,000 people, using the parameters given in Table 2. Without any intervention, containment will occur at Day 900 with 42,000 overall infections.

We observe some similarities as in the Rio de Janeiro State analysis. An increase in human recovery time delays the containment and increases the total number of infections (Figure 5a). Similarly, the incubation period for humans has little effect on the overall disease trend (Figure 5c). On the other hand, the incubation time for the vectors significantly affects the overall number of infections while exhibiting the same outbreak pattern (Figure 5d).

We explore alternative hosts and their impact on the pandemic. Increasing the number of alternative hosts from 20,000 to 100,000 (Figure 5b) will slow the outbreak and reduce the total number of infections by 83.3%. This suggests that introducing alternative hosts has potential in mitigating an outbreak. However, it is possible that these hosts can get infected with ZIKV and spread it to humans via other means. The ecological and disease impact cannot be fully assessed at this point, thus rendering it an unsafe venue for ZIKV disease containment.

We assume each strategy is implemented on Day 100 after the outbreak. Figure 6 reports containment effectiveness for each strategy. Reducing bite rate by 20% (Figure 6a) and increasing vector mortality by 20% (Figure 6b) have similar effectiveness in containing the disease spread (Day 500 vs Day 300) and with comparable overall infection (12,000 vs 11,000). Both strategies are very effective – reducing infections by over 70%.

Reducing larvae carrying capacity has a significant impact on the outbreak (Figure 6c). In particular, over 40% reduction in infection is observed when 20% reduction in carrying capacity is in effect. On the other hand, reducing vector population is not as effective in containing the ZIKV outbreak (Figure 6d). Nonetheless, reducing vector population does influence the pandemic course: when the vector population is reduced by 20%, it postpones the onset of the outbreak by approximately 100 days, and reduces the total infection by 9.5%. These results are different from the findings in the Rio de Janeiro State analysis where reducing the carrying capacity of larvae proves to be not very effective.

Figure 7 shows the combination strategies that can be implemented in practice in a cost-effective manner. Figure 7b shows the marginal gain in infection control through reduction in vector population.
Figure 5. Varying recovery time, alternative hosts, and incubation time for human and vector and their effect on overall infection and containment.
Discussion
This work demonstrates the key importance of “digital disease surveillance” in response to waves of epidemics including ZIKV, Dengue, Ebola and cholera. Specifically, we develop a compartmental model with a vector-host structure utilizing logistic growth for human population and a dynamic vector population describing its full life cycle. We analyze the development of ZIKV outbreak and explore key parameters that influence epidemic severity. We study the effect of ‘average’ recovery time (of human) from infection, the number of alternative hosts, and the ZIKV incubation time in both vectors and humans. The analyses are performed on two populations with real data from Rio de Janeiro State. Our results reveal that the outbreak is sensitive to the first three parameters while human incubation period has only marginal impact.

We examine four containment strategies: reducing biting rate, increasing vector mortality rate, reducing larvae carrying capacity, and reducing vector population. All four strategies are effective (in varying degrees) in containing the outbreak and reducing the overall infection. While findings for reducing biting rate and increasing vector mortality rate are similar for both populations (both are superior to other strategies), results for the larvae carrying capacity and reduction in vector population differ. These findings suggest environmental and demographic information should be considered when determining proper containment strategies. Our study also shows that reducing vector population early on can help delay the onset of the outbreak.

Combination strategies are both promising in practice and cost-effective in achieving early containment. The model provides a decision support framework for policy makers to estimate the cost-effectiveness for each prevention
measure. Public health departments should select a strategy portfolio compatible with their local environment and regional demographics. In addition, the public should be educated and informed of ZIKV status. Population behaviors (protecting themselves by reducing biting, cleaning water ponds to rid of larvae, etc.) have demonstrably significant impact on containing and mitigating the outbreak.

The multiple components involving the dynamics of human and vector populations in this model allow flexibility in characterizing disease spread and performing strategic analysis. However, obtaining/determining all essential input parameters for this to be practical may prove difficult. The model can be simplified without diminishing its quality and rigor in disease dynamics and infection/containment outcome prediction\(^5\). Such a model may be desirable when detailed input data are not readily available.

For example, the natural human population can be assumed to be constant throughout the outbreak since the magnitude of natural population growth is relatively small within the pandemic period. The mosquito population lifecycle can also be condensed to a simple birth-death process. In Lee et al 2016\(^5\), we propose a simplified model with constant human and mosquito population to study the ZIKV outbreak. The study predicts the disease trends and characterizes containment strategies for Brazil and Puerto Rico. The model has accurately predicted ZIKV outbreak trends for these two countries. It predicts the number of infectious pregnant women in Puerto Rico within 5% of the actual documented numbers (as of July 2016)\(^5\). Although the simplified model is statistically useful and usable, additional components and parameters presented herein should help decision makers better understand the details of ZIKV outbreaks. It may also provide them with more analytic-based options in deciding the proper containment strategies when such input data are available.

Modern information technologies which are capable of tracking disease spread and vector activities play an indispensable role in supporting disease modeling and in mitigating and containing epidemiology outbreaks. Informatics applications such as social media and internet-based participatory surveillance systems have already demonstrated their capability in helping us understand and prevent disease outbreaks like Dengue and influenza\(^2,5\). Software and hardware such as ArcGIS platform and mobile DNA sequencing lab are also designed to help track the current ZIKV outbreak. These technologies capture valuable inputs for our model, which is very flexible and can be easily adapted in real-time as real data feeds arrive. With the support of these technologies, the model parameters can be estimated accurately and updated on a regular basis. The model output can then be used to predict the status of the outbreak in a timely manner. It also reflects any changes in the environment, policy, and human and social behavior. Conversely, the model can also serve as a guideline of the application of these decision technologies such that they are deployed and utilized optimally by policy makers. Informatics can help us understand the underlying mechanism of the pandemic spread. It also enables decision and policy makers to understand/update the planning and response strategy in real time to maximize the containment effect while minimizing the overall infection / mortality and the associated costs and health burden.

There are limitations to our study. We assume a homogeneous human population in the model, which may not hold in practice. The possibility of contacting mosquitoes and the chance of getting infected from an infectious vector may differ among individuals and the environment. This can be overcome by creating additional compartments for each homogeneous subgroup. We have incorporated this subgroup concept in our vaccine prioritization analysis with great success\(^5\). However, this requires even more parameter estimation and may lead to over-fitting. In addition, our conclusions are strongly influenced by results from a particular demographic and geographic profile. Local decision makers should input proper (best estimate) parameters into the model in order to conduct accurate analyses.

The current model does not include any treatment or vaccination effects since they are currently under developed. With deeper understanding of ZIKV and the presence of medications and vaccines, these factors can also be incorporated under our modeling framework.

**Acknowledgement** The work is partially supported by a grant from the National Science Foundation and the Centers for Disease Control and Prevention.

**Reference:**
41. Erickson RA. Impact of climate change on the population dynamics of Aedes albopictus and dengue disease dynamics of dengue. [Master’s Thesis]. Texas Tech University; 2009.
Profiling Fast Healthcare Interoperability Resources (FHIR) of Family Health History based on the Clinical Element Models

Jaehoon Lee, PhD1, Nathan C. Hulse, PhD1,2, Grant M. Wood1
Thomas A. Oniki, PhD1, Stanley M. Huff, MD1,2
1Intermountain Healthcare, Murray, UT;
2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

In this study we developed a Fast Healthcare Interoperability Resources (FHIR) profile to support exchanging a full pedigree based family health history (FHH) information across multiple systems and applications used by clinicians, patients, and researchers. We used previously developed clinical element models (CEMs) that are capable of representing the FHH information, and derived essential data elements including attributes, constraints, and value sets. We analyzed gaps between the FHH CEM elements and existing FHIR resources. Based on the analysis, we developed a profile that consists of 1) FHIR resources for essential FHH data elements, 2) extensions for additional elements that were not covered by the resources, and 3) a structured definition to integrate patient and family member information in a FHIR message. We implemented the profile using an open-source based FHIR framework and validated it using patient-entered FHH data that was captured through a locally developed FHH tool.

Introduction

FHH has been recognized as a powerful tool for diagnosis and risk assessment in clinical genetics, aiming to identify risk of a specific gene disorder for the purpose of ensuring appropriate primary care and specialist intervention. However, utilization of FHH information has been limited by the lack of tools implemented in the clinical workflow to perform the required analysis. Unlike other clinical data, FHH includes not just patient data, but also their family members, including all relationships, their disease history, and age of disease onset or death. Traditionally clinicians have relied on interviews or questionnaires to collect narrative-only FHH information from patients, but due to the limitation of time in clinical settings, this method does not ensure sufficient information or data quality to complete the analysis. FHH is still mostly captured and stored in narrative text form, summarizing only essential or a minimum of information for risk assessment (e.g. family history of ovarian cancer on maternal side).

With support of computerized tools, improved electronic health record (EHR) systems are now capable of capturing and storing detailed FHH information in a structured way. Such tools have been implemented not only to be used by clinicians, but also outside clinical settings, as it ensures enough time and a systematic way to enter pedigree based FHH data by patients. Based on the ideal features of the FHH tools that have been suggested by the American Health Information Community, recently implemented tools provide a web-based access environment and visualization of family pedigree diagrams to connect with healthcare consumers, and motivate them to enter accurate FHH information.

In anticipation of increasing adoption of these tools, there have been efforts to develop a standardized data model of FHH for exchange across systems. The Health Level 7 (HL7) Version 3 implementation guide for family health history and pedigree interoperability is a well-known standard to represent FHH as an extensible markup language (XML) and has demonstrated its ability to be used for genetic risk assessment services across systems in a case study. However, it has been true that in general, HL7 Version 3 based standards haven’t been as widely adopted as Version 2, because Version 3 was thought to be difficult to develop, expensive, and required too much project time to develop.

FHIR is a next generation standard framework created by HL7 which has an implementation focused approach. FHIR provides an alternative to document-centric approaches by directly exposing discrete data elements, which compromise a FHIR resource, as services. Because the benefits of using FHIR are pre-built and standardized resources which are easily extensible, and the ability to combine these into profiles for easier implementation in development, a complex data model like FHH will be handled more efficiently using FHIR. Yet existing FHH resources do not cover detailed FHH data model such as full pedigree to support genetic risk assessment, and a few published extensions still have limitations.

In this manuscript, we used CEMs as a baseline data model to develop FHIR profile to address the gap. Elements of CEMs have features to be transformed to FHIR profiles in a straightforward manner, as the CEM strategy defines generic structures which specify the data type of an element, its permissible values or range, multiplicity of value
sets of the element, and permissible qualifying information, etc. \textsuperscript{13,14} In our previous study, we developed CEMs that can cover essential of FHH information; its key features are to capture full pedigree of FHH, linkages of family members to their patient records in EHR, and social history and health behaviors \textsuperscript{15}.

We selected FHH related resources by conducting a gap analysis between existing FHIR resources and FHH CEM elements, and added CEMs were not covered by the existing FHIR as extensions. We developed a structured definition that can contain all the selected resources and extensions as a patient’s full pedigree based FHH in one FHIR message. We used an open source based FHIR framework to implement the profile and to create messages using empirical FHH data of Intermountain Healthcare, and validated it by analyzing the gaps. We assume the profile to be used to run an external web service of genetic risk assessment.

**Background**

Intermountain Healthcare is a not-for-profit, community-oriented organization based on Salt Lake City in Utah with more than 37,000 employees, about 1,400 employed primary care and secondary care physicians, and provides healthcare services based on Utah and southeastern Idaho. It operates 22 hospitals and more than 185 clinics throughout the regions \textsuperscript{16}.

Intermountain Healthcare provides a web-based FHH collection tool named OurFamilyHealth, which was initially developed in 2009 and has made available through Intermountain’s patient web portal: MyHealth \textsuperscript{17}. The purpose of OurFamilyHealth is to provide patients the ability to enter FHH data by building a graphical family pedigree and assign disease and conditions to specific member in their history (see Figure 1). Patients are given a basic family tree at their first access and can add their family members to the tree with detailed demographic, disease histories, and health behavior information such as height, weight, and smoking status.

**Figure 1.** Screen shot: a graphically represented immediate family consists of a patient, father, mother, spouse, and two children in OurFamilyHealth

In our previous study, we introduced the principle and strategy of developing OurFamilyHealth, and also analyzed its early usage patterns \textsuperscript{18}. Although OurFamilyHealth was designed to capture and store essential FHH data elements based on the AHIC recommendation, yet it was not integrated with Intermountain’s EHR. As groundwork of the integration, we developed CEMs as baseline logic and data model to cover the essential of FHH information in OurFamilyHealth, to be eventually stored into our clinical data repository (CDR).

We currently have two use cases of utilizing the CEMs at Intermountain. First, OurFamilyHealth currently only collects patient’s FHH data but does not provide any decision support function to clinicians for patients. We have a
plan to enhance the application connected to an external risk assessment service such as HughesRiskService. To do so, OurFamilyHealth should be able to transform its FHH data into a standard message to run against the service, but also to interpret outcome of risk assessment from the service and present it to patients. The other motivation is integration of FHH data into our new EHR based on Cerner. Since OurFamilyHealth is a locally developed tool, it does not yet interface with Cerner system and a standard message can be used to exchange OurFamilyHealth data with Cerner’s CDR and applications. The data model and logic of the CEMs can be used in the both use cases.

Limitation of current model

HL7 FHIR builds on previous data format standards from HL7, like HL7 version 2.x and HL7 version 3.x, and is the most enhanced standard to exchange healthcare data. It is easier to implement because it uses a modern web-based suite of Application Programming Interface (API) technology, including a RESTful (Representational State Transfer) protocol, a choice of JavaScript Object Notation (JSON) or XML for data representation. FHIR is currently on second stage of draft standard for use (DSTU 2) and provides an alternative to document-centric approaches by directly exposing discrete data elements as services. For example, basic elements of healthcare like patients, admissions, diagnostic reports and medications can each be retrieved and manipulated via their own resource uniform resource identifiers (URIs).

A resource in FHIR is a small logically discrete unit of data exchange and has a known identity by which it can be addressed. Among FHIR resources, FamilyMemberHistory resource in the clinical category is to represent a simple structure used to capture an elementary family history for a particular family member. Figure 2 depicts a diagram of the data model of FamilyMemberHistory.

Figure 2. UML Diagram of FHIR Resource: FamilyMemberHistory (DSTU2)

Since FHH information consist of disease history of a patient and family members, ideally a combination of a Patient resource and a set of FamilyMemberHistory resources with appropriate linkages between resources can represent a FHH. The problem is that a FamilyMemberHistory only describes “a relationship of a family member to a patient” using the HL3 v3 relation codes, but it does not describe relationships between family members. For example, a FamilyMemberHistory resource instance with relationship code “PCOUSN” implies that a patient has a paternal cousin. But merely the relationship code does not examine who is parent of the cousin out of uncles and aunts, unless the patient has only one paternal uncle without aunts or one paternal aunt without uncle. Similarly, if a patient has multiple spouses and multiple children, a list of FamilyMemberHistory does not inform a child is from which of the spouses. As a result, we are unable to render a full pedigree from a set of FamilyMemberHistory.

FHH CEMs

The FHH CEMs are developed using Constraint Definition Language (CDL). CDLs were originally developed by GE Healthcare and are syntax for expression of constraint models that are key features of CEMs, with extensions including additions of new constraints to the modeling language, formalization of the model specialization capabilities, and a concrete reference object model. Based on our previous project, we have 126 CDLs to cover FHH including full pedigree, disease and conditions, and social history, such as smoking status, alcohol consumption status, and physical activity. Table 1 describes the key CDLs and their attributes.
### Table 1. FHH CDL

<table>
<thead>
<tr>
<th>Category / data element</th>
<th>CDL</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient</strong></td>
<td><strong>Identifier</strong></td>
<td>IREF</td>
</tr>
<tr>
<td></td>
<td><strong>Name</strong></td>
<td>ST</td>
</tr>
<tr>
<td></td>
<td><strong>Gender</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Adoption</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Race</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Ethnicity</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Living status</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Date of birth</strong></td>
<td>TS</td>
</tr>
<tr>
<td></td>
<td><strong>Date of death</strong></td>
<td>TS</td>
</tr>
<tr>
<td></td>
<td><strong>Age</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>Decease age</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>Multiple birth</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Cause of death</strong></td>
<td>CD</td>
</tr>
<tr>
<td><strong>Family relationship</strong></td>
<td><strong>Father</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Mother</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Children</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Marital status</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Consanguinity</strong></td>
<td>CD</td>
</tr>
<tr>
<td><strong>Disease and condition</strong></td>
<td><strong>Disease name</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Age of onset</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>Cause of death</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong></td>
<td>ST</td>
</tr>
<tr>
<td><strong>Health behavior</strong></td>
<td><strong>Height</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>Weight</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>BMI</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>International travel</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Physical exercise</strong></td>
<td>PQ</td>
</tr>
<tr>
<td></td>
<td><strong>Tobacco use</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Alcohol use</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Drug abuse</strong></td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td><strong>Note</strong></td>
<td>ST</td>
</tr>
</tbody>
</table>

* Data type: IREF (Item reference), PQ (Physical quantity), ST (String), CD (code), TS (Coded simple), INT (Integer)

**Method**

A FHIR profile is a set of base resources and is to author and publish a customized and more specific resource definition, by specifying a set of constraints and/or extensions on the base resource. Therefore concrete FHIR resources can express their conformance to a specific profile, and a FHIR server can to programmatically validate a given resource against the associated profile definition.
Our first step of profiling is to identify the base resources covering essential information of FHH, and then find extensions to cover rest of the information. We conducted a gap analysis between the FHH CDLs and current FHIR resources. The FHH CDLs are categorized by four meaningfully related groups: patient, family relationship, disease and condition, and health behavior. For every data elements in each category, we investigated if there is a match of FHIR resources, and then compared details with value sets in data elements in the two domains.

**Patient**

In a family pedigree represented based on the FHH CEMs, all the persons in the tree are recognized as patients and they are flat. However, we can also specify one person in the tree as a proband: a term used in genetic counseling to refer a patient describing his/her FHH to a counselor. In this case, a proband is a node locates in the center of a family pedigree, and relationships of all family members may be described as a set of “relationship to proband” (e.g. paternal father of patient). Data elements in this category cover mainly patient’s demographic information such as name, age, sex, and race / ethnicity, etc. Table 2 presents gap analysis between the data elements of proband category.

Table 2. Patient category: gap analysis

<table>
<thead>
<tr>
<th>CDL</th>
<th>FHIR resource</th>
<th>Gap analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>PatientExternalIdentifier</td>
<td>FamilyMemberHistory.identifier</td>
<td>EMPI* is used internally at Intermountain Healthcare. If a message is being used to share FHH between systems, system wide (or nation wide) master person index should be used to replace EMPI. If a message is being used to run risk assessment service and will not be stored anywhere, EMPI should be de-identified.</td>
</tr>
<tr>
<td>GivenName / FamilyName</td>
<td>FamilyMemberHistory.name</td>
<td>Exact match</td>
</tr>
<tr>
<td>AdministrativeGender</td>
<td>FamilyMemberHistory.gender</td>
<td>Exact match</td>
</tr>
<tr>
<td>AdoptedInd</td>
<td>FamilyMemberHistory.extension</td>
<td>HL7 relationship code is able to represent an adopted relationship of a family member, but it does not represent adopted by which family.</td>
</tr>
<tr>
<td>AdministrativeRace</td>
<td>FamilyMemberHistory.extension</td>
<td>Patient resource does not cover race, it may be added as an extension with value sets: White, Black/Afro American, American Indian, Asian Indian /Alaska native, Chinese, Filipino, Japanese, Korean, Vietnamese, Native Hawaiian, Guamanian / Chamorro, Samoan, Pacific islander, Other, Unknown</td>
</tr>
<tr>
<td>AdministrativeEthnicGroup</td>
<td>FamilyMemberHistory.extension</td>
<td>Patient resource does not cover race, it may be added as an extension with value sets: Non-Hispanic, Hispanic/Latino, Ashkenazi Jewish, Other, Unknown</td>
</tr>
<tr>
<td>DeceasedInd</td>
<td>FamilyMemberHistory.deceased</td>
<td>deceasedBoolean may be used.</td>
</tr>
<tr>
<td>BirthDate</td>
<td>FamilyMemberHistory.born</td>
<td>BirthDate has a qualifier of approximate flag. This should be added as an extension.</td>
</tr>
<tr>
<td>Age</td>
<td>FamilyMemberHistory.age</td>
<td>Exact match</td>
</tr>
<tr>
<td>DeceasedDate</td>
<td>FamilyMemberHistory.deceased</td>
<td>deceasedDate may be used.</td>
</tr>
<tr>
<td>MultipleBirthInd</td>
<td>Observation.category</td>
<td>This presents that a person belongs to one of multiple birth category. Observation.category = social-history</td>
</tr>
<tr>
<td>MultipleBirthType</td>
<td>Observation.valueString</td>
<td>Twin, Triplet, Quadruplet, Quintuplet, Sextuplet</td>
</tr>
<tr>
<td>CauseOfDeath</td>
<td>FamilyMemberHistory.extension</td>
<td></td>
</tr>
</tbody>
</table>

* EMPI: Enterprise Master Person Index

**Family relationship**

The family relationship structure in FHH CDLs is described by the Family Panel contains three items: parent1, parent2, and child (See Figure 3), which means a unit family basically consists of two parents and one child. The model of the three items is Patient, which implies every family member in the Family Panel is considered as patients and also identified by PatientExternalIdentifier that may be linked to patient records in EHR systems. AdministrativeMaritalStatus and Consanguinity represent marital information of parent1 and parent2. If a person has relationships with multiple spouses, Family Panels may be needed as many as the number of the spouses. For example, if person A has married twice with spouses B and C, and has had children D and E from with spouse, the relationship will be presented with two Family Panels consist of the following items: {A,B,D} and {A,C,E}.
Figure 3. Family Panel in CEM presented as CDL.

Table 3. Family relationship category: gap analysis

<table>
<thead>
<tr>
<th>CDL</th>
<th>FHIR resource</th>
<th>Gap analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent1</td>
<td>FamilyMember.extension</td>
<td>Internal Id of a parent in a FHIR message</td>
</tr>
<tr>
<td>Parent2</td>
<td>FamilyMember.extension</td>
<td>Internal Id of a parent in a FHIR message</td>
</tr>
<tr>
<td>Child</td>
<td>FamilyMember.identifier</td>
<td>Internal Id of a child</td>
</tr>
<tr>
<td>Consanguinity</td>
<td>FamilyMember.extension</td>
<td>Consanguinity means that a father and a mother have another relationship than marriage.</td>
</tr>
<tr>
<td>AdministrariveMarital Status</td>
<td>FamilyMember.extension</td>
<td>This is to represent marital status for parent1 and parent2 at the time of data entry. Value sets: {married, divorced, partner, separated}</td>
</tr>
</tbody>
</table>

Disease and condition

Disease and condition history of proband and family members is essential information to assess risk of genetic diseases. Not all disease and condition concepts are captured; the FHH CEMs may bind consumer-facing disease concepts or family history. These concepts are mapped to SNOMED concepts. When a FHIR message is created, local disease code should be transformed to a SNOMED code to ensure interoperability with external systems.

Table 4. Disease / condition: gap analysis

<table>
<thead>
<tr>
<th>CDL</th>
<th>FHIR resource</th>
<th>Gap analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinicalAssert</td>
<td>FamilyMemberHistory.condition.text</td>
<td>SNOMED code</td>
</tr>
<tr>
<td>AgeOfOnset</td>
<td>FamilyMemberHistory.condition.onsetQuantity</td>
<td>Age of onset as number</td>
</tr>
<tr>
<td>Cause of death</td>
<td>FamilyMemberHistory.condition.outcome</td>
<td>This is to specify a disease is cause of death or not.</td>
</tr>
<tr>
<td>Note</td>
<td>FamilyMemberHistory.condition.note</td>
<td>This is free text to describe details of disease history.</td>
</tr>
</tbody>
</table>

Health behavior

The FHH CEM also contains data elements of health behaviors, which are potentially related to genetic risk assessment as environmental factors (e.g. smoking history of family with lung cancer). We mapped Observation resource to represent health behaviors as they may cover a diagnosis, monitor progress, baselines and patterns, and demographic characteristics.

Table 5. Health behavior: gap analysis

<table>
<thead>
<tr>
<th>CDL</th>
<th>FHIR resource</th>
<th>Gap analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>HeightMeas</td>
<td>Observation.category Observation.code Observation.value</td>
<td>Observation.category = “vital sign”</td>
</tr>
<tr>
<td>BodyWeightMeas</td>
<td>Observation.category Observation.code Observation.value</td>
<td>Observation.category = “vital sign”</td>
</tr>
<tr>
<td>BodyMassIndexMeas</td>
<td>Observation.category Observation.code Observation.value</td>
<td>BMI can be optional if in case height and weight are known. Otherwise it can be an independent measure.</td>
</tr>
</tbody>
</table>
From the gap analysis in four categories, we identified five types of gaps to address between the two models.

- **Exact match**: concepts of data elements are same with their data types and value sets (e.g. first / last name, date of birth).
- **Mapping required**: race and ethnicity in the two domains have same concepts but different value sets and need to be mapped. For example, OurFamilyHealth captures “Ashkenazi Jews” particularly in ethnicity, as they statistically have more BRCA1 and BRCA2 genes that imply higher risk of breast, ovarian, and prostate cancer [25], however HL7 ethnicity code does not support it.
- **Reasoning required**: e.g. inferring a family member’s adoption status in his/her family from a HL7 relationship code of the person and proband.
- **Type conversion**: age in years can be calculated from date of birth. However date of birth can be only estimated as a range from age in years.
- **No match**: existing FHIR resources or HL7 value sets do not cover consanguinity. It should be added as a new Observation or extension.

**FHIR structure**

In addition to selecting the base resources and extensions, we built a model of structure to integrate them in one FHIR message. We used two infrastructure resources: Bundle and List. Bundle refers to a container for a collection of resources, particularly to gather a collection of resources into a single instance with containing context. List is to contain a set of information summarized from a list of other resources and is a flat, possibly ordered, collection of records. Resources supported by the List resource can be homogenous – consisting of only one type of resource or heterogeneous – containing a variety of resources. Here we used List to contain homogenous resources such as Condition and FamilyMemberHistory. Using the features of Bundle and List, a structure of a FHH FHIR message can be formed as Figure 3.

In the structure, a family pedigree of a patient can have 1) a HeaderList that contains information of message header such as identifier of a FHIR message, message created (or sent) date, language code, etc., 2) a PatientBundle that contains a patient’s conditions and observations, and 3) a FamilyMemberHistoryList that contains multiple FamilyMemberHistory resources and their conditions and observations. All resources in the structure have resource references to a patient that indicate who is proband in a family pedigree. In addition, all Observations and Conditions of family members have subject references that indicate whom in the tree this clinical data belongs to.
Implementation
We implemented the proposed profile using HAPI framework; an open source based library for creating and validating FHIR messages. HAPI is pure Java based, and licensed under the business-friendly Apache Software License, version 2.0. Developers can create a FHIR message reusing Java objects, methods, and value sets provided by the libraries. Using these components, we developed a composer to generate a FHH FHIR message based on the profile we created.

To validate the FHH profile with HAPI, we used empirical patient data from OurFamilyHealth. FHH data in the database of OurFamilyHealth is stored in form of XML. We developed a parser to extract the XMLs from the database and decompose them as Java objects at the level of the FHH data elements. We automated parsing each patient’s FHH XML into individual data elements then feed them into the parser to generate FHIR XMLs. In the process we de-identified the data prior to creating messages by eliminating the tags of personal identification including names, birthdates of patients and families. As a result, we successfully created 4594 XMLs of FHH FHIR message (See example message in Figure 4).

Conclusion and future research
Our approach of deriving the profile of FHH FHIR through gap analysis based on the CEMs was plausible. We successfully implemented the profile and created FHIR messages using an open source based FHIR framework and empirical patient data.

We have limitations in our study and future works to address them. First, although our FHH CEMs were developed to be a general data model for FHH, they were developed based on the local data model of OurFamilyHealth. In addition we used patient data from OurFamilyHealth for validating the profile. Because of this the profile should be more validated in outside Intermountain settings and use cases. Secondly, although we assumed that the profile would adopt the feature of the FHH CEMs that can link family members to patient records in EHRs, we have not validated the idea. In our use case, we assumed that all the patients are independent and don’t have any links each other.
There are ongoing efforts to validate and implement the created profile in real use cases. We are working with the team of HughesRiskApp to connect OurFamilyHealth to their risk assessment service through FHIR. Currently we are validating the profile based on the historical patient data to see it can properly be consumed and interpreted by the service. In the future we plan to run the service real-time. For example, at the end of a session that a patient enters FHH into the application, he/she will be presented risk analysis results given the information entered.

**Figure 4.** An XML of FHH FHIR message created based on HAPI. Two extension tags represent two parents of a patient (left), and their internal person Ids (Id = 1, 2) match person Ids of FamilyMemberHistory (right).

We are also working with Healthcare Services Platform Consortium (HSPC): a provider-driven organization of healthcare institutes, IT vendors, systems integrators, and venture firms, providing a collaborative platform to achieve the gold standard of semantic interoperability of healthcare applications. HSPC is working on general profiling of FHIR based on CEMs, and our FHH profile will be part of the process eventually to be validated, publicized, and standardized.

**References**


9. HL7 Version 3 Implementation Guide: Family History/Pedigree Interoperability, Release 1


Toward Medical Documentation That Enhances Situational Awareness and Learning

Leslie A. Lenert, M.D., M.S
Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA

Abstract

The purpose of writing medical notes in a computer system goes beyond documentation for medical-legal purposes or billing. The structure of documentation is a checklist that serves as a cognitive aid and a potential index to retrieve information for learning from the record. For the past 50 years, one of the primary organizing structures for physicians’ clinical documentation have been the SOAP note (Subjective, Objective, Assessment, Plan). The cognitive check list is well-suited to differential diagnosis but may not support detection of changes in systems and/or learning from cases. We describe an alternative cognitive checklist called the OODA Loop (Observe, Orient, Decide, Act). Through incorporation of projections of anticipated course events with and without treatment and by making “Decisions” an explicit category of documentation in the medical record in the context of a variable temporal cycle for observations, OODA may enhance opportunities to learn from clinical care.

Introduction

Physicians and other healthcare providers write notes for a variety of purposes, including documentation of care, communication with other team members, and providing evidence of services for billing. There are a growing number of applications that help physicians write notes both for efficiency and to produce coded data as a by-product. Whether application generated, dictated, typed or (increasingly rarely) hand written. There is a substantial question as to whom physicians are composing notes for (billers, themselves, or other providers) and many notes by medical professionals go largely unread by others. (1) There is also a growing concern about the use of cut and paste actions and templates short circuiting clinical cognition (2) and about resident practices that separate note composition from clinical workflows pre-empting thinking during note writing. (3)

We believe that when physicians compose notes, they should be thinking about the patient and the task of composition should shape both the processes of care and the ultimate conclusions they reach about diagnoses.(4) Moreover, note writing should provide an opportunity to learn from a case both as it evolves and on retrospective review. This paper reviews predominant cognitive models that shape providers’ thinking during the process of documentation, and proposes an alternative model designed to promote situational awareness and learning from experience.

Background

Larry Weed is the father of structured recording of medical information to facilitate reasoning. His theories were first expounded almost 50 years ago in the seminal paper, “Medical Records that Guide and Teach.” (4) At the core of Weed’s theories is the use of checklists for clinical cognition that encourage medical providers to think systematically about their patients. A widely adopted structural innovation described by Weed—the Subjective, Objective, Assessment and Plan (SOAP) framework—reminds clinicians of specific tasks while providing a framework for evaluation of information and provides a cognitive framework for reasoning activities.

The “Subjective” header is a cue for documentation of data that comes from patients’ experiences. By labeling this data “subjective”, it encourages clinicians to reason using an objectivist view and to divide the world into the “subjective” experiences of patients (history of illness, symptoms experienced, pain and anxiety and other features as elicited by the clinician or other healthcare providers), and the more “objective” measures of the clinician (physical examination findings, diagnostic testing and radiological examinations) that govern the processing of subjective reports of illness into medical diagnoses. Donnelly critiques this division of the world into subjective and objective based on the perspective of the physician: “The fact that we cannot know the facts becomes obscured if one accepts the subjective-objective distinction in SOAP.” (8)

The Assessment section documents the synthesis of “subjective” and “objective” evidence to arrive at a diagnosis. The Assessment section details the differential diagnosis process and may include some risk benefit tradeoffs in decision making. The Plan section details the needs for additional testing and consultation with other clinicians to address the patients’ illnesses, in addition to the steps being taken to treat the patient.
Examining this framework, what is the operational task that it supports? This author would assert that task is capture and consideration of all relevant data with a view toward differential diagnosis. SOAP is probably not the actual framework that clinicians use in diagnostic reasoning (9), but it has probably served well as a guide to diagnostic reasoning for a variety of reasons. The framework is hypothetico-deductive one that is focused on the application of the scientific method to patient care and the disproof of competing hypotheses. This gives the approach a certain face value (that may be diminishing in light of “big data” methods of empiric assessment and treatment). In addition to being “scientific”, the approach is fundamentally Bayesian: the prior probability of different diseases is updated by elements personal history (the subjective). Physical exam findings and laboratory tests, further update the posterior probability, ruling in or out different diagnoses. The Assessment follows all this data, and thus is allows the clinician to review relevant data before reaching a conclusion. Thus, if differential diagnosis was the only task that providers performed, SOAP might be the ideal cognitive framework.

But what if the situation is more dynamic? Could hypothetico-deductive reasoning lead providers astray by anchoring them on past observations and conclusions? It is likely that more common task in medical environments than differential diagnosis of an unknown problem(s) is the active management of an evolving, known medical problem over time and the differentiation of change in a disease from acquisition of a new one. In many clinical situations, evidence changes over time, requiring providers to reconsider diagnoses and treatments or risk error. (10) The SOAP model does not explicitly integrates time into its cognitive framework and thus has an important gap.

This has not gone unnoticed. Nurse clinicians have proposed extensions to SOAP designed to address its temporal “blind spot.” For example, the SOAP, SOAPIE, and SOAPIER models add “Intervention”, “Evaluation” and “Revision” headers and, as a result, creates categories for linking a note to overall progress on treatment of the problem. (11) The “E” and “R” are important refinements to SOAP. “E” adds an explicit reminder to assess how well a plan worked. “R” creates explicit expectations that plans need to change over time.

Physician have also proposed changes to their SOAP notes to make it easier follow a problem over time. Proposed changes have focused frames designed to help providers look backward across prior notes may detect changes in a patient’s condition or failure of a medical plan to achieve its intended effects. To support these concepts, Meyers and Miller suggest adding an “orientation” step to SOAP (O-SOAP) based on retrospective review of records. (12, 13) Gensinger and Fowler suggest re-arrangement of SOAP as ASOP (14), focusing on Assessment at the beginning to provide more temporal orientation. An ASOP structured note would bear some similarity to a cognitive frame widely used handoffs between providers called SBAR (Situation, Background, Assessment, Recommendations) (15), with Assessments in ASOP playing a similar role to Situations in SBAR. While putting Assessment first has some advantages, it might also lead to faulty decision making, if performed in order of the acronym, by forcing conclusions prior to review of all the data, voiding the positive cognitive check list effects of SOAP.

Unfortunately, none of the proposed enhancements of the SOAP cognitive framework have gained wide acceptance as a way to document progress notes. One potential reason for the “stickiness” of the SOAP model is that it is cognitively “right-sized.” As Miller described in his seminal 1956 paper (16), there are limits on working memory capacity of the brain. Cowan Miller’s work arguing that the magic number of items is four (17). It is an interesting coincidence that SOAP has four steps—as does SBAR, which is also widely adopted (18-22), and that extensions to SOAP that add additional tasks have not received as wide acceptance. Alternatively, SOAP extensions may not offer sufficient advances as to warrant their adoption based on Rogers Theory of Diffusion. (23)

Where could we look to improve upon SOAP, borrowing from other fields, while also recognizing the importance of prior authors insights into the limitations of SOAP? Electronic health records systems (EHRs) offer the potential of re interpreting the SOAP approach to focus on the evolution of problems over time. As described by Meyers’ et al. (12), the problem with the SOAP is that “it does not recognize the crucial effect that the last assessment and plan and the returned plan results have upon the next visit.” Prototypes developed by Meyers et al. (13) and Bosen (24) , illustrate how EHRs could bring elements of SOAP notes related to specific problems over time together in context to improve reasoning. However, there might also be value in changing the frame all together.

An Alternative Framework for Reasoning

The discipline of human factors research provides alternative directions on how to reorganize note writing. The failure of a clinician to recognize change in manifestations of a hypothesized disease over time, or to recognize unexplained systems, might be characterized as a lack of situational awareness.(10) Endsley defines situational awareness as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.” (25) For example, situational awareness might require that physicians retain unexplained observations in their working memory and to document these in their notes (for example, the fever that doesn’t resolve or the serum sodium that is too low), in addition to
documenting their diagnoses and their rationale. Or, situational awareness might require one discipline to be aware of the findings of another. For example, a physician might need to read a nurse's note to be aware of a patient's changes in cognition. Situational awareness might also include subtle observations of interactions between the patient and family members that give clues about how care for the patient occurs at how.

Similar to SOAP, there are cognitive models that outline steps that can help novices and experienced users maintain situational awareness and learn in rapidly changing environments. One such model, which was initially developed by Boyd to help Air Force pilots outthink their opponents in aerial combat but is now widely applied within the U.S. military, is the Observe Orient Decide Act (OODA) model.

OODA is an action-oriented decision making model is explicitly applied in a repeating loop. Completion of one cycle prompts the next. There are no fixed time intervals for OODA cycles—the rate of cycling depends on how fast circumstances are changing. The goal of the OODA Loop is “think” and “act” faster and better than the enemy (in this case the disease(s) afflicting the patient), and thereby utilize available resources to maximum effect. (26) This is an important concept that is rarely addressed in medical reasoning: unless the pace of medical assessments and interventions outpaces the disease, the patient will, at best, experience morbidity or at worst, mortality (27). Bad outcomes occur when medical treatments are not adjusted fast enough, based on current hypotheses about what the illness is, to outpace progression of the illness.

In addition, OODA loops emphasize recognition of mismatch between predictions and experiences and, also, automation of responses in specific settings, once a problem has been recognized. OODA and other models that formally integrate a predictive step enhance learning through observation. They force a projection on the learner and periodic (re) assessments to test that projection with reality.

The OODA Loop is a proven technique to speed cognitive adaptation to changing circumstances. It has been widely adopted within U.S. military doctrine and the success of operations such as the first Gulf War, attributed, at least in part, to its appropriate implementation. (26, 28) OODA loops have been successfully utilized in the field of medicine (29, 30) but not as a template for documentation or teaching reasoning. The specific elements of the OODA framework will be discussed in turn in following paragraphs.

The Observation step encompasses both subjective and objective elements found in the SOAP note avoiding the subjective-fallacy. The decision maker notes what facts are available at a particular time. Observations might be general or specific to a designated problem. They might come from more than one observer; for example, a nurse or a caretaker. Observations may reflect patient history or be abstracted from other elements of a clinical record such as laboratory and radiological tests. Observations are time stamped, reflecting the state of the world at the time of its examination by the reasoner. This is critical, because in retrospective review of decision making, for the purposes of learning, a learner would need to distinguish an error due to a lack of availability of key information (a failure of observation) from a failure to integrate that information into a decision process.

The next step in the OODA model is Orientation, which puts Observations in context and presents integrative hypotheses for causation or explanation (e.g., diagnoses and social contexts). To quote Boyd, “orientation is an interactive process of many-sided implicit cross-referencing projections, empathies, correlations, and rejections that is shaped by and shapes the interplay of genetic heritage (!), cultural tradition, previous experiences, and unfolding circumstances.” A frequent cause of decision-making error in complex, evolving environments is cognitive tunneling, which is a form of “expectancy-driven” thinking bias. (31) An explicit goal of the orientation step is the prevention of this mental tunnel vision and the mitigation of confirmation bias—that is, an excessive, self-convincing focus on a few details thought to be pertinent. Orientation includes noting which observations are missing and which are unexplained.

Boyd stressed iterative re-orientation as an essential means to correct oneself from perceptual “mismatches” with reality, as it “unfold[s] in an irregular, disorderly, unpredictable manner.” (32) This effortful act of sense making (33) would call upon the physician to apply the wholeness of his or her being, including his or her previous experience and innate talents, medico-ethical heritage, a holistic contextual view of his patient, and even his or her feelings in addition relevant guidelines and resources available in the environment in order to integrate across observations and problems to create a synthesis on the current patient state. Orientation might include updating probabilities of different potential diseases in the differential diagnosis. It could also include an assessment of the effects of treatment, but should always include the non-medical context of care. Problem list generation and, perhaps more importantly, problem list maintenance, are part of the Orientation step.

Endsley (a contemporary of Boyd, whose work includes best practices for safety in aircraft cockpit environments and has been adopted by medical safety researchers) expanded the initial concepts underlying the OODA Loop by proposing that situational awareness should include projection (prediction) of the effects of both action and inaction. (25, 34) In current documentation practices, providers rarely predict what will happen if they do nothing or expected the effects of a prescribed treatment. This is critical for subsequent comparison to observed
effects to ascertain if the cognitive model of the provider is correct. For example, predicting the effect of prescription of a blood pressure medicine (e.g., drop in systolic blood pressure of 10 mm/hg by the next visit, a decrease in potassium of 0.2 ug/mL after infusion, or an increase in uric acid concentration of 0.5 ug/mL). Documentation of the expected effect of actions (and non-actions) is important in assessing the correctness of decisions and the adequacy of actions. Understanding of the gap between predictions and observations is critical to learning from experience. To echo Boyd, who also made projection a major theme of his work, if your predictions were incorrect, your earlier observations (perceptions) or orientation (integration of information) were mismatched with reality. One must –(re) observe and (re) orient, quickly until predictions match reality. Or, in other words, one must learn.

The third step in the OODA model is to make and record decisions. Decisions include both interpretations of observations through the lens of orientating (e.g., diagnoses) and risk-benefit trade-offs that underlie Actions. Decisions, like Observations and Orienting hypotheses, should be time stamped. (3) Analyzing the effect of decisions provides the opportunity to assess and refine decision-making skills—to allow trainees to learn from their experiences and to see, potentially, what others decided to do in similar circumstances. Quality control efforts could benefit from individual and team performance reviews of decisions made during episodes of care. Tracing longitudinal decision pathways permits individual and team reflection and learning. It may also facilitate the identification of novel approaches, which could be leveraged across a healthcare organization in pertinent care settings, or inform guideline improvement efforts. For example, natural language process approaches might be applied to identify all the decision points in a case and bring those together for review. In SOAP formatted notes, medical decisions are often implicit in documentation, embedded in Assessments and Plans; there is no easy way to call these out of a health record and re-examine a provider’s reasoning vis-à-vis the state of evidence at any particular time and his or her rationale for decision making.

Advanced versions of OODA include implicit guidance and control operations as an alternative to explicit decision-making. In note writing and clinical care, this might take the form of instructions for “what if” cases that are expected. For example, a standing order for use of acetaminophen with a fever over 39 C, followed by obtaining a blood culture. These bits of “compiled” knowledge of how to respond to a patient’s illness shorten the time to implementation of orientating insights and potential enable quicker control of progression of disease.

The final element in an OODA cycle is to specify Actions. The OODA framework specifies Actions rather than the SOAP note’s Plans: this is because it is one thing to plan and another to act. While the difference is semantics to some degree (Webster’s defines a “plan” as “a detailed formulation of a program of action”), specifying actions removes ambiguity. For example, a list of actions might clarify which team members have responsibility for execution of the different actions that comprise “the plan” (including the patient). Action specifications should include predictions of their effects, potentially aided by predictive analytics models. Observations in subsequent cycles should address the effectiveness of implementation of proposed actions.

The OODA Loop has the advantage of being applied in a cyclic fashion, clarifying thinking over time, with the most recent Observations challenging Predictions and Actions from the previous period. Iterative diagnostic reasoning helps prevent errors by encouraging the review of assumptions of prior decisions and aiding to assess the degree to which the operational model fits observed effects. (35) A failure to reconsider initial diagnosis in the face of mounting evidence supporting an alternative is a frequent cause of error. (36) The Decide step, in a temporal frame, allows revisiting of the choice, model, and actions, and a reconsideration of assumptions. In contrast to SOAP, OODA prompts the documenter to refine his or her thinking based on how prior actions correspond to current observations. (See Figure 1.) It prompts the note writer to contemplate, “What was I thinking yesterday and was it correct? How should I think differently today? What is important to focus on at this juncture? What is relevant? What warrants action? What will happen if I do nothing?” (37) While the OODA cycle includes prediction in two of the steps, we would argue that prediction is not an independent action. Rather one cannot/should not complete the tasks of Orientation or Action selection without the context provided by a prediction of the outcomes of inaction and action.
As clinicians increasingly “cut and paste” content from prior patient encounters into notes, (38) the decision support opportunities provided by note writing are in jeopardy. It is time we (once again) recognize note writing as a cognitive task that helps clinicians think better and thereby avoid errors. But, in order to better reflect what we know about clinical cognition after 50 years, perhaps it is time to advance the mindset that guides notes. There is probably not a single cognitive frame that is valid for all types of notes. The cognitive frame of note writing should match the predominant task in management of the patient. The POMR/SOAP format is well adapted to the initial medical encounter. In an initial evaluation of a patient, the sequential evaluation of symptoms, signs and evidence to formulate a differential diagnosis may be the most important task. Further, a reductionist approach, focused on identifying specific problems and solutions, may be highly adaptive. But, in the subsequent management of patients, particularly in dynamic environments or the ongoing management of patient conditions, it may be more important for the cognitive frame of the note to aid the author in maintaining situational awareness, and in detecting change and errors in diagnosis. In this setting, integrating lessons from OODA-style thinking may be preferable. An OODA-style note would also improve communications by explicitly documenting a clinician’s reasoning and decisions, much of which is not transparent in SOAP notes.

What would an OODA note look like in an electronic medical records system? To better understand this, working with a student in my laboratory, we adapted from the initial visit of a patient with colorectal cancer to the OODA format. In Figure 2a is the description of the patient visit in a more traditional SOAP and problem oriented format. Figure 2b is the description in the OODA format. The OODA note is not comprehensive in order to illustrate its differences.

While the first note focused on the process of coming to the diagnosis of colorectal cancer and the evidence for an against this diagnosis, the OODA equivalent focused more on developing an action plan that both executable and acceptable to that patient. The consequences of inaction are also made explicit. The OODA note is more conceptual and combines subjective and objective data elements in one category and orientating information that places the patient in a human context. The explicit predictions with and without treatment help assess both the
correctness of the diagnosis and the impact of care. In a SOAP note, it is relatively easy for the note to focus the note on the disease (colorectal cancer) rather than the patient. The note, a documentation of the findings of the physician, goes into great detail on the history suggestive of cancer, the findings that confirmed it, and emphasized the amount of time it took for the physician to perform the careful history and examination required to make the diagnosis. It is an excellent logbook entry (15) of the care of this patient. But, beyond careful documentation, what did does the log do? In this case, the diagnosis was virtually certain (and would be absolutely certain with obvious diagnostic tests.) What was uncertain that needed to be managed? It was the patient’s response to learning the diagnosis and the plan for controlling his symptoms. OODA may help focus documentation and as a result clinical thinking on the uncertain—on the parts of care that are not obvious—that need attention and future care. The differences are illustrative and should not be considered the target for OODA implementation rather than the definitive instance.

**Subjective** - A 65 year old male presenting to address the following issue(s):
Patient presents with a 2 month history of weight loss, blood in stool, and diarrhea. Also has moderate to severe LLQ pain for which he is taking very large amounts of acetaminophen (3-4 grams per day) and had been taking excessive amounts of naproxen (750 mg BID). Pain is in the LLQ and RLQ. Keeps him up at night in addition to diarrhea. Was using some sort of fiber-bulking agent colon “cleaner” for about 2 months (March to April) prior to onset of diarrhea.

**Problem 1: Loss of Weight**
Eating a regular diet with good appetite but has had diarrhea and a 20 lb. weight loss over the past 3 months. Easts everything without pain or discomfort. No fever chills or night sweats. Moderate LLQ and some RLQ pain, mostly at night.

**Problem 2: Blood in Stool**
Ongoing over past several months. Small amounts of blood in stool. No bleeding between bowel movements.

**Problem 3: Chronic Diarrhea**
Floating smelling stools with blood on occasion 4 to 20 times per day without fever or chills

**Past medical history**
No problems

**Family history**
Father with brain tumor
Brother with some sort of heart trouble

**Medications**
OTC naprosyn and tylenol for pain as described

**Social History Narrative**
- Married
- Smokes between 5 and 20 cigarettes per day
- No alcohol or other substances
- Delayed access to care until he was eligible for Medicare despite symptoms because of costs/lack of insurance

**Review of Systems**
Constitutional: Positive for fatigue and unexpected weight change.
Gastrointestinal: Positive for abdominal pain, diarrhea, blood in stool, anal bleeding and rectal pain.
Musculoskeletal: Positive for back pain
Skin: Positive for rash
Psychiatric/Behavioral: Positive for sleep disturbance.

**Objective**
BP 106/65 | Pulse 77 | Temp 98.4°F | Resp 16 | Ht 5’11” | Wt 146 lb 14.4 oz | BMI 20.49 kg/m2 | SpO2 92%
Constitutional: He is oriented to person, place, and time. No distress observed. Thin white male.

Normal HEENT examination
Normal Pulmonary examination
Normal Cardiovascular examination except for S4
Normal Abdominal examination, including liver and spleen size
Genitourinary: Penis normal. Gualac positive stool. Rectal:
Exophytic hardish rough mass about 3 cm beyond anus. Felt like it was very prominent and at least 2X2 cm. Very painful to touch (deep pain).
Unable to get around for complete rectal exam. Opposite wall smooth and without lesion. Unable to palpate prostate due to mass. Pink tinged heme positive fluid on glove after exam
Lymphadenopathy: Right side:
inguinal adenopathy present. No supraclavicular and no epitrochlear adenopathy present. Left side: inguinal adenopathy present. No supraclavicular and no epitrochlear adenopathy present. Small shotty nodes.
Extremity examination: Normal
Neurological examination: Normal

**Assessment/Plan**
Rectal mass: rule out rectal carcinoma
Chronic diarrhea/to infection
Weight loss due to malignancy or malabsorption
CT scan with contrast of abdomen and pelvis
Hydrocortone for pain, diarrhea-caution about maximum safe dose of acetaminophen
Labs: CBC, Chemistries, LFT’s, PT, stool culture, UA
Referral to colorectal surgeons to evaluate patient
Follow-up on Friday

Figure 2a. Description of an initial visit of a patient with colorectal cancer in the traditional format.
Other Medical Applications of OODA documentation

In addition to using OODA as a documentation template, there are many other potential applications that could take data from an OODA framework and make it more available to clinicians. If data in OODA format were coded and if appropriate meta data were available, then it might be possible to re-present notes created in the OODA format in the more traditional SOAP format. In fact, there is probably not a strong reason to align note creation frameworks supporting cognition with the framework used for display for a viewer. Informatics tools could allow readers to view notes in the templates most aligned to their own thinking using metadata to order and organize the display.

OODA may be an important tool in medical education. Notes labeled in an OODA format allows a different type of chart review of a case more akin to “replay” of the clinical state of thinking. At each point where a decision was recorded, the learner could review the observed data, the orientations of the decision maker, the projects and the decisions. In current EHRs, review of a case occurs primarily through the retrospective lens of the discharge summary. While many authors write discharge summaries as historical records or “captain’s logs”, this is probably not the highest use of these documents, which should be focused on care coordination post discharge.(15) The ability to record decision points and to review decisions over time might be one of the most important advances afforded by OODA approaches.

OODA may also have a roll in coordinating care across disciplines. One way it could do this is by providing an interdisciplinary “blackboard” for recording observations, orienting facts and hypotheses, decisions, and actions that are transdisciplinary. One reason that physicians, nurses and other providers might not view each others documentation as often as they should (1), might be the orientation of that documentation around SOAP models targeted at diagnosis specific to their own discipline. Medicine (and even subspecialties of medicine), nursing, dentistry and other disciplines all make diagnoses of that often may be of limited relevance to other disciplines. But what if as each discipline were writing their note, they could mark up observations, orientations, etc. as being important for the

Figure 2b. Same note in OODA format.
situational awareness of other providers. This might not take re-entry of data—rather in a coded data environment, one might mark up a fact for sharing. The OODA model might be very good for organizing data sharing as a neutral framework. A transdisciplinary blackboard might also allow providers to identify facts that are unexplained or request help from other disciplines in confirming predictions.

Conclusions

While the SOAP model has served practitioners well for over 50 years, the nature of the tasks that physicians perform has evolved over that time, its limitations are well known. The purpose of this article is to introduce OODA, a tested cognitive framework that has potential lessons for how doctors should write notes to improve the clarity of their thinking and to make their reasoning more transparent so that the writer and subsequent readers of notes can learn from care. OODA reflects the evolution of SOAP over time but also introduces new concepts related to temporality and explicit decision cycles for clinical care. This paper does not address the issues of how to link OODA notes together or the potential medical-legal issues of having every decision open for review in a medical record. These are complex issues; however, pragmatic issues, should not be allowed to thwart strategies to improve patient care. Small improvements in cognitive aids for note writing used many times a day by hundreds of thousands of physicians might have large impacts. Research is needed to identify the optimal framework for cognitive support during note writing. OODA is an example of a proven framework that may have important implications for medicine and which could potentially serve as a replacement for SOAP.

Acknowledgements.

This paper is dedicated to a former student who wishes to remain anonymous but deserves credit for introducing me to the concepts of advocated by Boyd and his OODA Loop. While I must honor his wishes for privacy, I also wish to express my heartfelt thanks for teaching-his-teacher, and for his insights and enthusiasm.

Funding Statement

This work was supported by NLM Training Grant No. T15LM007124.

References

13. Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 1956;63::81–97.
Predictive Analytics to Support Real-Time Management in Pathology Facilities

Lysanne Lessard, PhD1,2, Wojtek Michalowski, PhD1, Wei Chen Li, BSc1, Daniel Amyot, PhD1, Fawaz Halwani, MD, PhD, FRCPC3,4, Diponkar Banerjee, MBCHB, FRCPC, PhD3,4
1University of Ottawa, Ottawa, Ontario, Canada; 2Institut de Recherche de l’Hôpital Montfort, Ottawa, Ontario, Canada; 3Eastern Ontario Regional Laboratory Association and The Ottawa Hospital, Ontario, Canada; 4University of Ottawa, Faculty of Medicine, Department of Pathology and Laboratory Medicine

Abstract

Predictive analytics can provide valuable support to the effective management of pathology facilities. The introduction of new tests and technologies in anatomical pathology will increase the volume of specimens to be processed, as well as the complexity of pathology processes. In order for predictive analytics to address managerial challenges associated with the volume and complexity increases, it is important to pinpoint the areas where pathology managers would most benefit from predictive capabilities. We illustrate common issues in managing pathology facilities with an analysis of the surgical specimen process at the Department of Pathology and Laboratory Medicine (DPLM) at The Ottawa Hospital, which processes all surgical specimens for the Eastern Ontario Regional Laboratory Association. We then show how predictive analytics could be used to support management. Our proposed approach can be generalized beyond the DPLM, contributing to a more effective management of pathology facilities and in turn to quicker clinical diagnoses.

Introduction

The introduction of new tests and technologies in anatomical pathology is bringing at once new opportunities and challenges for diagnostic pathologists. For example, the onset of personalized medicine requires that new methods for the detection of oncogenic pathways in tumors be integrated in routine diagnostic pathology [1]. This shows an increasingly important role for pathology in management of surgical patients; however, it also implies a much higher volume of surgical specimen to be diagnosed, along with pathology processes more complex than what is currently the norm in pathology facilities. Given that timely pathology diagnostics are required to provide optimal patient management, changes in terms of the clinical complexity of pathology processes create challenges at the managerial level. Indeed, clinical and operations managers of these facilities need to ensure that pathology processes continue to run safely, efficiently and effectively despite their growing size and complexity.

The use of advanced information technologies and analytics in pathology – otherwise known as pathology informatics – has grown in response to these challenges [2]. In healthcare, predictive analytics have now matured sufficiently to be used in addressing major issues, such as preventing early post-discharge readmissions, facilitating patient engagement, and detecting insurance and billing fraud [3]. These techniques now extend visual analytics systems such as dashboards that allow users to access relevant data through graphical and tabular displays, drill down in the data and form hypothesis about root causes. However, dashboards often lack the capability of predicting events [4]. In this paper, we discuss how advanced predictive analytics can be integrated at key points in pathology processes in order to facilitate workload, throughput planning, and intervention by managers of pathology facilities. For the purpose of this paper, we define predictive analytics rather broadly as a set of techniques such as data mining, statistics, modeling, and machine learning to analyze current data and make predictions about future events. In that sense, we accept that predictive analytics is based on a fundamental assumption that patterns observed in the past are sufficiently stable for predicting future behavior. While this assumption is sometimes contested, its limitations can be mitigated by creating predictive models that can adjust to changes in past patterns through a feedback loop. However, it is important to stress that despite these limitations, the use of predictive analytics in healthcare is being recognized as a reliable approach to improving outcomes, enhancing patients’ experience, and reducing the costs of delivery of health services [5].

We illustrate the potential contributions of predictive analytics using the problem and solution domains at the Department of Pathology and Laboratory Medicine (DPLM) at The Ottawa Hospital (TOH), Ottawa, Ontario,
Canada. TOH is a teaching hospital affiliated with the University of Ottawa. The DPML houses grossing, histology and cytology laboratories, and processes all pathology specimens for the Eastern Ontario Regional Laboratory Association (EORLA), a newly established association of all the laboratory and pathology departments of Eastern Ontario that currently includes facilities from eight hospitals in the region. The rapid growth of volume of surgical and cytology specimens resulting from this centralized structure has created many challenges in insuring smooth workflow and processing of cases and diagnostic reports in a timely fashion. A number of research projects are underway to address these challenges and provide solutions that support decision-making for managers of the DPML, including the development of optimized pathologists’ scheduling models [6], a real-time visual dashboard tracking the inventory to be processed by the DPML [7], and predictive analytics algorithms for workload and throughput planning. We focus here on the latter.

The remainder of the paper is organized as follows. First, we describe the pathology processes currently implemented at the DPML, and identify three key points within the surgical specimen process workflow where predictive analytics would generate the greatest benefits. We then present the systems architecture currently in place at the DPML, and situate the use of predictive analytics within this architecture. We elaborate on the use of predictive analytics that is the most appropriate for each point within the specimen process flow. We conclude with future work and contributions.

Pathology processes at the DPML

We briefly present the pathology process for surgical specimens implemented at the DPML at TOH. The process revolves around pathology-specific entities being managed at the DPML: cases, specimens, blocks, and slides. A case groups all specimens received from one patient-physician encounter. A specimen is one or more tissue fragments removed from an organ or specific site. A block is a container of one or multiple tissue fragments from the same specimen. A slide is a thin slice of tissue cut from the block and stained with a specific set of reagents. These elements thus form a hierarchy of the inventory to be processed by the DPML.

The first step in the DPML process is the accessioning of specimens, where specimens and clinical orders from physicians arrive and are given a uniquely identifying case code. Each specimen is also associated with a pathology subcode that determines how many blocks and slides will need to be created for this specimen. The second step is the grossing of specimens, where blocks are created from the tissue and automatically processed in a manner suitable for creating slides later on. In the third step, blocks are sent to the Histology laboratory, where each one is manually embedded in paraffin and then sliced and stained as required by the pathology subcode attached to it. In the fourth step, called dispatch, finished slides are checked to ensure that they are matched with the correct specimen and case. They are then organized by case and assigned to pathologists. The last step is the diagnosis, where pathologists receive complete cases to examine. During this diagnosis process, a pathologist may order additional slides or blocks. The process ends when the pathologist makes a definite diagnosis for a case and writes a report.

Figure 1 shows a simplified business process model developed for surgical specimens at the DPML. As can be seen in the model, the process requires interactions among four key entities (Grossing lab, Histology lab, Immuno lab, and Pathologist) in order to be completed. This is not a fully linear process, since requests for additional slides or tests can originate from pathologists when they start to interpret slides, preventing the diagnostic process to be completed until these requests have been fulfilled and quality controlled. Key challenges in ensuring the timely completion of diagnostic reports are (i) the variability in the amount and type of cases being received at any point in time; (ii) the optimal allocation of cases to individual pathologists given the variability in cases as well as the variability in their availability and sub-specialties; and (iii) the variability in levels of complexity across similar cases, which can only be ascertained once the diagnostic process has been initiated. These challenges highlight the potential benefits of predictive analytics at three specific point in the surgical specimen flow shown in Figure 1: predicting the number of cases and case types that will be received by the pathology facility, predicting the workload that will be generated from this volume, and predicting the actual diagnostic throughput of pathologists based on case volume and complexity.
The DPLM currently uses the commercial laboratory information system PowerPath® by Sunquest [8] to record and track individual cases. PowerJ, a dashboard application developed in-house, pulls data from this system in order to monitor key phases in pathology processes at the DPLM, namely: i) number of pending and grossed cases; ii) number of pending and cut blocks; and, iii) number of pending and routed slides. The application also records and presents individual pathologists’ workload, and overflow slides by sub-specialty. A business intelligence (BI) tool providing a fine-grained overview of the performance of the DPLM in terms of cases that are within or that exceed predetermined processing targets has been developed and is planned to be deployed in the near future. The latter combines data from PowerPath® with business rules that specify target processing times per specimen types and process steps. The BI tool supports monitoring, analysis, and reporting. It does not, however, offer the ability to predict future events. Therefore, a predictive analytics capability needs to provide this added level of support. Figure 2 relates these current and upcoming capabilities as a proposed system architecture for the DPLM.

**Current and upcoming information systems at the DPLM**

The DPLM currently uses the commercial laboratory information system PowerPath® by Sunquest [8] to record and track individual cases. PowerJ, a dashboard application developed in-house, pulls data from this system in order to monitor key phases in pathology processes at the DPLM, namely: i) number of pending and grossed cases; ii) number of pending and cut blocks; and, iii) number of pending and routed slides. The application also records and presents individual pathologists’ workload, and overflow slides by sub-specialty. A business intelligence (BI) tool providing a fine-grained overview of the performance of the DPLM in terms of cases that are within or that exceed predetermined processing targets has been developed and is planned to be deployed in the near future. The latter combines data from PowerPath® with business rules that specify target processing times per specimen types and process steps. The BI tool supports monitoring, analysis, and reporting. It does not, however, offer the ability to predict future events. Therefore, a predictive analytics capability needs to provide this added level of support. Figure 2 relates these current and upcoming capabilities as a proposed system architecture for the DPLM.
Figure 2. Proposed systems architecture at the DPLM

Facilitating effective management in pathology through predictive analytics

The transition of pathology diagnosis into a context of digital pathology where slide analysis is conducted using computer-based tools represents a significant step forward in the pathological diagnosis process. This transition also creates new opportunities for the production, capture and use of large volumes of accurate data. In the DPLM, data capture is currently confined to selected stages of the process illustrated in Figure 1. Namely, data are collected at the start and completion of some, but not all, of the steps. With the advent of digital pathology, additional events can be captured, allowing for example to determine the amount of time a pathologist actually spends on each slide during the diagnosis step. Once laboratory information systems have been adapted to capture these new data, they can be leveraged to provide pathology facilities managers with a more accurate picture of their facilities’ operations. Our current project with the DPLM is designed to take advantage of such emerging opportunity.

The real-time use of operational data has the potential to increase the quality of diagnosis and to reduce the costs of the diagnostic process, contributing both to an increase in facility performance and effective management [9]. However, the contribution of pathology informatics to effective management has received less attention in the academic community; given the growing complexity of pathology processes and the resulting challenges to the management of pathology facilities, applying predictive analytics to support managerial decision-making is timely and relevant. This is the focus of this paper and we explain here the benefits of using predictive analytics at three key points within the DPLM surgical specimens process flow: predicting the number of incoming cases by type (and associated diagnostic complexity), predicting the workload of the pathologists, and finally predicting the throughput of the facility. Next, we describe our proposed solutions for each of these areas.
Predicting the number of incoming cases per type

The DPLM operates in a constantly changing and seemingly unpredictable environment. The arrival of surgical cases is driven by a surgical schedule, the allocation of operating room blocks to surgical specialties (thus determining types of the surgeries), and the complexity of patients’ cases. In most hospitals, there is a disconnect between the information infrastructure designed to support patient management and surgical services (patient registration, EHR, surgical information system) and the one supporting pathology services. Hence, pathologists and managers of pathology facilities are typically not aware of the surgeons’ schedules and changes to block scheduling of the surgeries. This situation makes it difficult to correctly allocate resources within a pathology facility, for example in terms of scheduling the optimal number of laboratory technicians on a given day.

In our research, a long-term objective is to address this disconnect by developing an application that will capture relevant data about surgeries and store this information in a relational database (bottom of Figure 2) to be accessible by the pathology BI tools. While this will give access to reliable data on the number of surgical cases to be processed on any given day, it will not give a complete understanding of the level of complexity of incoming cases. However, historical data about past surgeries and types of incoming cases will allow us to develop predictive models using statistical techniques. Namely, we will use case complexity as the dependent variable and develop logistic regression models to predict the behavior of this variable in the future. Simultaneously, we will be mining pathology data in order to develop associations between case complexity and pathology specific attributes such as number of slides, stains used, etc. In this stage of our work, we expect to rely on unsupervised data mining methods such as association rule learning [10].

Predicting pathologists’ workload

The need to measure workload in a pathology facility is critical for management in order for appropriate service levels to be maintained. The Canadian Association of Pathologists advocated an approach to workload measurement that is based on the L4E system [11]. The L4E system was developed from survey data collected from 27 pathology centres; the results suggest that a full-time pathologist working in a facility such as the DPLM should generate about 7560 L4E equivalents annually. While this measure can be used to evaluate individual pathologists’ performance, it is not very useful for predicting the workload of pathologists working in a teaching hospital; indeed, such predictions need to take into account their academic, teaching, research, and administrative responsibilities – which can account for up to 75% of their time -, as well as dimensions such as each pathologist’s sub-specialties and the impact of each case’s complexity on the time needed for diagnosis.

Our research thus takes a different approach; specifically, the association rules for case complexity used to process historical surgical will allow us to translate purely surgical information into data annotated with the complexity of pathological diagnosis. In turn, we will be able to use the annotated data as an input to a stochastic optimization model for predicting pathologists’ workload taking into account their availability, sub-specialties, prioritization of the cases, and sessional variations. Having such predictive power should allow managers of pathology facilities to better plan human resource allocation by sub-specialties, to develop flexible work schedules (including “floating” resources when available), and to prepare contingencies for unexpected events.

Predicting throughput

As Figure 1 shows, there are a number of potential bottlenecks in a pathology process, from the time cases are received to the time the pathology report is submitted. In order to help manage the process, and take full advantage of digital pathology, it is important to capture fine-grained data at each input and output point in each process step. This may however require a change in practices and equipment. For example, at the DPLM, data about each case, specimen, block, or slide is usually captured when the element arrives at a workstation, and at the time that the work is completed, but not necessarily at a fine-grained level. For example, when a pathologist receives slides to be interpreted for a case, data is captured at beginning of the step (by scanning one of the slide), and again at the end of the process, once the report is finalized; however, events such as interruptions do not generate data. Hence, data about the duration of some steps does not differentiate between active work and waiting time.

Capturing data at the times when an element leaves one station and moves to another would give more precise information from which to predict future throughput. Moreover, while processing times in some instances are fixed and cannot be really manipulated (thus act as fixed parameters), there are some that can be influenced by adding...
additional resources to processing. The resulting real time throughput prediction model could monitor the actual volume of cases that are in a process, that are waiting for processing, that are being diagnosed, or that are waiting for a diagnosis, hence representing a general performance measure for a pathology facility. Going beyond performance measurement, simulation models could simulate different strategies of dealing with changes in incoming cases and workload parameters, and be associated with the throughput prediction model. Once a model would detect deviation from predicted throughput (for a given context described as day of a month, etc.), it would automatically consult one of the scenarios and apply it (or present it to a manager) in order for the detected anomaly to be addressed.

Implementing the solution

Deployment of the proposed predictive analytics solution will be part of a multi-layer architecture that clearly separates data from a host of analytics tools, and user interface. The logical, hence application-neutral architecture is illustrated on Figure 2. The advantage of such a multi-layer architecture includes the ability to combine varied components, for example a hospital-wide data infrastructure in the data tier with a business intelligence application specifically developed for a pathology facility. While the proposed architecture can be implemented in a number of different (and customized) manners depending on the data infrastructure and suite of analytics solutions ultimately chosen, the DPLM implementation will involve using one of the industry predictive engines (such as those developed by IBM, SAS, or SAP). We are planning for the data tier to include stream data coming from, for example, laboratory information systems such as PowerPath® and historical pre-processed data, for example surgical block schedules updated occasionally. This tier will also include the association rules for determining case complexity and the business rules describing the processing of the specimens in case of the unexpected events (developed as a result of the simulation described earlier). Pragmatic issues regarding the integration of these data streams, for example variations in data structures, will have to be addressed.

Application and presentation tiers will act as an interface between data and end-users. The application tier will include a host of modeling services together with a set of BI applications. The presentation tier is where the reports, alerts, etc. are to be configured. We will use ontology-driven design to create a repository of presentation widgets for easier re-use and customization. Finally, a client tier will implement a web portal that will represent a “one point entry” to all the functions our system offers to the DPLM managers. Access to the portal will be governed by role-based principles, with different levels of granularity of information being accessible to different types of the end-users. We are currently working on finalizing the general architecture of the system, on identifying the relevant software tools, and on eliciting a final set of system requirements from different groups of potential end-users and other stakeholders.

Preliminary results of a user study conducted to evaluate a prototype of the business intelligence layer illustrated in Figure 2 shows that pathology managers find the proposed solution relevant to their work and plan to use the final system to support daily decision-making when it will be implemented in DPLM. Moreover, the preliminary version of the proposed solution had a positive impact on DPLM operations, allowing for example for more effective scheduling of pathology assistants and adjustments to pathologists’ schedules [7]. Given these positive results, it is expected that a fully implemented solution will increase managerial efficiency and effectiveness at the DPLM, allowing the facility to face new challenges in terms of the volume and complexity of cases to be processed.

Conclusion

At a time of growing complexity in pathology processes, it is of great importance that the management of pathology facilities be adequately supported in order to ensure that processing times and quality of provided services remain within acceptable standards. Indeed, unresolved managerial issues could trump the benefits of the new tests and technologies that are currently being introduced to support new patient management paradigms such as personalized medicine. In this paper, we have shown that predictive analytics could usefully be applied in a pathology facility to address issues of variability in case volume and complexity that make it difficult to adequately plan resource allocation over short and long time horizons. Specifically, we proposed the use of techniques such as logistic regression models, data mining, and simulation models to predict the amount of incoming cases per type, complexity, workload requirements, and facility throughput. These techniques could provide novel and timely solutions to the managers of pathology facilities beyond the DPLM, contributing to a quicker processing of pathology specimens and in turn of clinical diagnoses.
References

Comparing lagged linear correlation, lagged regression, Granger causality, and vector autoregression for uncovering associations in EHR data

Matthew E. Levine, BA1, David J. Albers, Ph.D.1, George Hripcsak, M.D., M.S.1
1Department of Biomedical Informatics, Columbia University, New York, New York, USA

Abstract

Time series analysis methods have been shown to reveal clinical and biological associations in data collected in the electronic health record. We wish to develop reliable high-throughput methods for identifying adverse drug effects that are easy to implement and produce readily interpretable results. To move toward this goal, we used univariate and multivariate lagged regression models to investigate associations between twenty pairs of drug orders and laboratory measurements. Multivariate lagged regression models exhibited higher sensitivity and specificity than univariate lagged regression in the 20 examples, and incorporating autoregressive terms for labs and drugs produced more robust signals in cases of known associations among the 20 example pairings. Moreover, including inpatient admission terms in the model attenuated the signals for some cases of unlikely associations, demonstrating how multivariate lagged regression models’ explicit handling of context-based variables can provide a simple way to probe for health-care processes that confound analyses of EHR data.

Introduction

With the increasing collection and storage of patient electronic health data around the world comes a proportionally growing impetus to use that information to improve clinical care. These improvements can range from workflow and operations optimization to pharmacovigilance studies, but the central feature for effectively exploiting the electronic health record (EHR) is our ability to learn from the data collected. We hope to move towards reliable high-throughput methods for determining adverse drug effects that can be applied to large clinical data repositories, like that collected by Observational Health Data Sciences and Informatics (OHDSI), which contains over 600 million patient records [1].

Many research inquiries can be satisfied with simple determinations of whether a patient ever had a particular condition, and it is often sufficient to consider events that occur over relevant time windows with respect to a condition of interest [2]. However, it can be useful to consider methods with the potential to reveal fine temporal structure in EHR data, and recent advances in such methods have been applied to machine-learning approaches during phenotyping [3,4], pattern discovery [5–7], temporal abstraction over intervals [8], and dynamic Bayesian networks [9].

Many of these approaches to time-series analysis rely on assumptions of stationarity (roughly, having consistent mean and variance through a time window of interest) that are frequently broken by clinical data—this is to be expected, even desired, since the primary goal of medicine is to drive patients from problematic to healthy states. This issue is compounded by the simple fact that patients are sampled with greater frequency when they are ill [10]. In fact, it appears that clinicians sample patients at rates proportional to their health variability, effectively inducing stationarity by indexing the time series not by clock-time, but rather by mere measurement sequence with single units of time imposed between each measurement [11].

Our past work has revealed informative results about temporal processes in the EHR by applying lagged linear correlation to time series constructed using linear temporal interpolation and intra-patient normalization of clinical signout note and laboratory test data [12]. These results indicated temporal processes that were definitional (e.g. low potassium levels associated with hypokalemia), physiologic (e.g. a potassium-sparing diuretic preceded increases in potassium levels), or intentional (e.g. a potassium-sparing diuretic was ordered in patients with low potassium levels), and used clock-time as the lagged time variable. Similarly, time-delayed mutual information reveal lagged linear structure as well as nonlinear dynamical processes related to physiology [13,14] despite EHR-data complexities and homo- or heterogeneity among patient populations [11,15–17].

Our most recent efforts to characterize temporal processes in the EHR are motivated by our previous findings that 1) temporal clinical and physiologic processes can be described through lagged linear correlation of concepts extracted from signout notes and laboratory values [12], 2) time series data, under some clinical circumstances, are better parameterized by their raw sequence than their clock measurements [11], and 3) health-care process events such as inpatient admission are systematically correlated with concepts and laboratory values [18].

779
In this study, we used multivariate distributed lag models to incorporate additional context-related variables in lagged linear analysis of temporal processes to better characterize both intended and unintended physiologic effects of drugs. In order to broaden the applicability of the method, we designed a time series preparation methodology that can use drug-order records as inputs, which are readily available in more contexts than physician notes. In order to evaluate these methods, we applied them to twenty pairings of drugs and laboratory measurements. As part of optimizing time series construction methods, we also investigated the effects of two pre-processing steps: intra-patient normalization of laboratory tests and different data preparation strategies.

Because our goal is to minimize bias and confounding, we employed two techniques to minimize bias. We used a particular form of lagged regression, known as Granger causality [19], to assess the effect of one variable (drug) over another (laboratory measurement) beyond that accounted for by the target variable’s autocorrelation. We used an extension of Granger causality, vector autoregression [20], to also account for a third variable (inpatient admission) as an example of a health care process confounder.

Methods

Experimental design

We used the 27-year-old clinical data warehouse at NewYork-Presbyterian Hospital, which contains electronic health records for over 3 million patients, to examine pairwise relationships between drug order records and laboratory measurements. We considered five drugs—simvastatin, amphotericin B, spironolactone, and warfarin—and four laboratory tests (total creatine kinase (CK), creatinine, potassium, and hemoglobin), and a patient cohort was identified for each of the 20 drug-lab pairs in the experiment. We identified eight drug-lab pairs for which clinical evidence suggested significant physiologic associations (shown in Table 1); we did not find conclusive evidence for associations between the remaining 12 drug-lab pairs. Patients were included in a drug-lab cohort if they met the following criteria: 1) at least 2 of the laboratory measurements of interest on record, 2) at least 1 order for the drug of interest, and 3) more than 30 combined data points between laboratory measurements of interest and total drug orders (any drug). No attempts were made to remove or correct outliers.

Building a time series from clinical data

Laboratory measurements, drug orders, and inpatient admission events for each patient in each cohort were extracted from the clinical data warehouse. A piecewise-defined linear drug-lab timeline was constructed for each patient as described by Hripcsak et al. using linear temporal interpolation (see Figure 1 in Hripcsak et al. for an example) [12]. Laboratory values were continuous, and orders for the drug of interest were represented as 1 (present), whereas orders for other drugs were represented as 0 (absent). Although orders for other drugs do not necessarily indicate cessation of a drug of interest, such orders were treated as evidence of absence to avoid incorporating external domain knowledge about drug administration that might produce artifact associations. Inpatient admission timelines were defined with a 1 at the time of admission, and zeros at 24hrs before and after admission, effectively creating spikes at times of admission. Smoothness and differentiability of the admission spike are unimportant when using discretized convolutional approaches, and there are many ways of constructing the spike such that it has mass to contribute during the convolution. For every time point where there was a concept (lab, drug, or inpatient admission), the values of each other variable at that time point were interpolated as the clock-time weighted mean of the preceding and succeeding value of each respective variable (or as the closest measurement if there was no value on one side). Clock-time weighting is computed by weighting the 2 bordering values by their temporal distances from the time-point at which we interpolate. Thus, all concepts, whether from categorical or real-valued sources, took on rational values that were paired at each time point. Because each time-point typically has only 1 reported event, each time-tuple is comprised of 1 true data point and 2 interpolated values.

Pre-processing of time series data

Two types of pre-processing of interpolated time series were designed and evaluated. First, each patient’s time series of laboratory values were normalized to have 0 mean and 1 variance by subtracting the mean and dividing by standard deviation [12]. This operation removed inter-patient effects. Second, we replaced each interpolated value in the time series with its difference from the immediately preceding interpolated value, such that time series values represented changes in values. This was an important means of reducing dependence between lagged variables in our novel application of multivariate lagged regression models to interpolated clinical data. As such, we effectively considered 4 time series construction methods—1) no pre-processing, 2) intra-patient normalization, 3) differences, and 4) normalization and differences.
**Sequence time**

Although clock-time was used for weighted interpolations and pre-processing steps, it was discarded in favor of raw sequence time for subsequent lagged linear analyses due to recent examples that demonstrated greater information and greater stationarity in clinical time series data that are parameterized by their sequence [11]. Real-time may prove to be a more sensible choice in other circumstances, especially when data is stationary. All time intervals between interpolated, pre-processed values were set to unit 1 length, effectively converting from clock-time to sequence time.

**Univariate lagged linear regression (ULLR)**

We compute lagged linear regression coefficients, $\beta_\tau$, for the following distributed lag model [20], where $y_t$ represents a laboratory value at sequence time $t$ and $x$ represents the interpolated drug value at time $t-\tau$. This model performs the same computations as our previous lagged linear correlation experiment [12], but supplies a different statistic, namely the lagged drug coefficient $\beta_\tau$.

$$y_t = c + \beta_\tau x_{t-\tau} + \epsilon_\tau$$  

(1)

**Multivariate lagged regression (MLLR)**

We aim to leverage the relationship between the context of each data point and the variables they predict by including context-dependent factors in multivariate autoregressive models. In general, a multivariate distributed lag model for $L$ lags and $N$ variables (for which the $i^{th}$ variable is denoted $u_i$), can be used to define the lagged coefficient for each variable $u_i$ at each lag $\tau$ (denoted as $\beta_{u_i,\tau}$), and is written as [20]

$$y_t = c + \sum_{i=1}^{N} \sum_{\tau=1}^{L} \beta_{u_i,\tau} u_{t-\tau}^i + \epsilon$$  

(2)

where $\beta_{u_i,\tau}$ is the coefficient for lag $\tau$ of the variable $u_i$. Because the parameters of these models are estimated jointly, adding explanatory variables that are related to both the predicted variables and variables of interest can change the values of coefficients of interest. Concretely, we considered a simple multivariate lagged regression that only incorporates lagged drug values, and jointly estimates all lagged drug coefficients $\beta_\tau$, with $L=30$, according to the following model, which we refer to as the “multivariate lagged drug model”

$$y_t = c + \sum_{\tau=1}^{L} \beta_\tau x_{t-\tau} + \epsilon$$  

(3)

We then evaluated how adding lagged terms to represent previous laboratory values affects drug coefficients by fitting the following “autoregressive drug and lab” model with $L=30$; this is in the form of Granger causality [19].

$$y_t = c + \sum_{\tau=1}^{L} \beta_{y,\tau} y_{t-\tau} + \sum_{\tau=1}^{L} \beta_{x,\tau} x_{t-\tau} + \epsilon$$  

(4)

We also introduce an additional context variable, $z$, to represent the inpatient admission timeline, and fit a further augmented “autoregressive drug, lab, and context” model with $L=30$; this is in the form of vector autoregression [20].

$$y_t = c + \sum_{\tau=1}^{L} \beta_{y,\tau} y_{t-\tau} + \sum_{\tau=1}^{L} \beta_{x,\tau} x_{t-\tau} + \sum_{\tau=1}^{L} \beta_{z,\tau} z_{t-\tau} + \epsilon$$  

(5)

Intuitively, this model uses the last 30 interpolated laboratory values, the last 30 interpolated drug values, and the last 30 interpolated admission values from the constructed time series to predict a present measurement. This alignment of previous data is performed for each laboratory measurement, and is aggregated within each patient, then across patients, creating a matrix with 91 columns (90 explanatory values and 1 predicted value) and a length equivalent to the number of qualifying laboratory measurements in the cohort. We did not perform any feature selection procedures, such as Bayesian information criterion, as this was out of the scope of our case study for...
method comparison—employing such selection criteria is a key step in determining true Granger causality, which we did not attempt.

Bootstrap

After cohort identification, timeline construction, and pre-processing, we aggregated pairs of lagged interpolated values across patients to construct a sparse model matrix for each drug-lab pair that conformed to the specified dimensionality of each model. Coefficients for each model were estimated by performing sparse linear least squares regression with Cholesky factorization from the MatrixModels package in R [21]. One hundred iterations of a bootstrap were performed for each matrix by sampling patients with replacement [22] in order to obtain empirical estimates of statistical significance. Coefficient estimates were labeled as statistically significant if zero was not included in their 95% Confidence Interval (CI) as computed by the bootstrap, which was defined as

\[
\left( \mathbb{E}[^\beta] - 1.96 \sqrt{\text{var}(\beta)}, \mathbb{E}[^\beta] + 1.96 \sqrt{\text{var}(\beta)} \right)
\]

In our evaluation, we focus on the estimates of lagged drug coefficients, and evaluate the effect of additional variables not by examining their coefficients directly, but rather by evaluating how their presence affected the drug coefficients.

Results

Intra-patient normalization in univariate lagged linear regression

Univariate lagged linear regression (ULLR) with intra-patient normalization was performed for each of the 20 drug-lab pairs of interest, eight of which we hypothesized, based on clinical literature, to have a significant directional (i.e. increasing or decreasing) effect. The number of significantly positive and negative lagged drug coefficients are compared with results that could be expected from the literature in Table 1. Analysis of normalized data with ULLR detected 4 out of 8 expected signals correctly with appropriate directionality (3 positive relationships and 1 negative), and reported all other cases to have statistically significant relationships. While this demonstrates real statistical correlations between variables, it does not necessarily implicate a physiologic association. The many biases in EHR data can be misleading when drawing statistical conclusions, so it is important to find ways of systematically focusing the analysis to reveal only the effects of interest—in this case, ones rooted in physiology.

In Figure 1, we show that univariate LLR analysis of normalized data reveals clinically characterized trends, such as amphotericin B’s tendency to decrease potassium levels and increase creatinine. It shows an overall trend linking spironolactone to increases in creatinine levels (a known phenomenon), but also finds a statistically significant negative relationship at lag of 1 in sequence time. Figure 1 also indicates a significant association between simvastatin and increases in hemoglobin levels, for which we do not have a particular biological interpretation. This result is characteristic of the significant signals detected by ULLR in the 12 drug-lab pairs for which we did not expect a physiologic association. The many biases in EHR data can be misleading when drawing statistical conclusions, so it is important to find ways of systematically focusing the analysis to reveal only the effects of interest—in this case, ones rooted in physiology.

In Figure 1, we show that univariate LLR analysis of normalized data reveals clinically characterized trends, such as amphotericin B’s tendency to decrease potassium levels and increase creatinine. It shows an overall trend linking spironolactone to increases in creatinine levels (a known phenomenon), but also finds a statistically significant negative relationship at lag of 1 in sequence time. Figure 1 also indicates a significant association between simvastatin and increases in hemoglobin levels, for which we do not have a particular biological interpretation. This result is characteristic of the significant signals detected by ULLR in the 12 drug-lab pairs for which we did not expect a physiologic association. These results suggest that univariate LLR analysis, like other analytic approaches to clinical data, is vulnerable to health-care process effects in EHR data. For example, the short-term negative relationship of spironolactone and could be attributed to a treatment pattern in which patients first prescribed the drug are likely to be sick (possibly with high creatinine) and subsequently improve due to treatment. Creatinine elevation due to the drug, then, would be on a longer time scale than creatinine-lowering treatments. Table 1 shows that the multivariate models remove this effect, likely by jointly considering previous orders of the drug.
Adding autoregressive terms: multivariate lagged linear regression

We adopted a multivariate LLR model in order to address the confounding effects of health care process when tasked with detecting true physiologic effects of drugs. We first considered a model that estimates all lagged drug coefficients jointly (equation 3, “multivariate lagged drug model”) with the intent of incorporating drug timeline history into each prediction. We also evaluated an augmented version of this model that adds autoregressive terms of previous laboratory values as well as drug orders (equation 4, “autoregressive drug and lab model”)—this formulation is very similar to the autoregressive model used in Granger causality, although we do not perform model selection to choose the number of lags, nor do we address potential unit root issues as is typical in Granger causal analysis [19]. We used intra-patient normalization and applied the differences pre-processing step described earlier (replacing values of the time series with their difference from the previous value) in all multivariate LLR analyses to reduce dependence between interpolated values (independence of lagged variables is an important requirement in multiple regression).

The multivariate lagged drug model showed statistically significant relationships with correct directionality between 6 of the 8 drug-lab pairs with known biological activity and rejected 3 of the 12 uncharacterized pairings. The autoregressive model of drug and lab histories detected the same 6 of the 8 known pairs, and also rejected 3 uncharacterized pairings. Moreover, in cases of true associations, drug and lab autoregression revealed, on average, more significant coefficients with greater magnitude. This finding suggests that the multivariate drug and lab model will have better sensitivity than the lagged drug model without sacrificing specificity.

Figure 2 shows how univariate LLR, multivariate LLR with drugs, and multivariate LLR with drugs and labs describe two cases: 1) ibuprofen predicting creatinine levels, and 2) simvastatin predicting changes in hemoglobin. It is known that ibuprofen can cause acute renal failure, for which high creatinine is a common symptom, and indeed we see that multivariate LLR (and not ULLR) predicts the correct directionality of the effect, with larger and more significant coefficients produced when using autoregressive terms of lab and drug values. An effect of simvastatin on hemoglobin levels, however, was not supported by evidence from our literature search, suggesting that apparent effects are attributable to non-biological phenomena. Figure 2b shows that the MLLR drug model predicts a significant negative effect, whereas the lab and drug model indicates no significant relationships. This implies that previous hemoglobin measurements helped to explain future drops in hemoglobin (i.e. this model may account for the number of anemic patients on simvastatin). The univariate model, however, showed a positive effect during early time points, indicating something more likely related to health care processes that were accounted for by the multivariate models through joint parameter estimation.

The performance of multivariate LLR methods was dependent on both intra-patient normalization and taking differences of the time series data sets. Omitting both pre-processing steps resulted in 0 correctly identified signals for the drug-only model, and 1 correctly identified signal for the drug and lab model. Using normalization without differences allowed the drug and lab model to detect 1 additional signal (2 in total), while using differences without normalization restored performance closer to levels seen with both pre-processing steps. The drug-only model detected the same 6 of 8 hypothesized drug-lab pairs, and had 10 apparent false positives. The lab and drug model significantly underperformed without both differences and normalization, detecting half the number of expected associations when used without normalization.

Furthermore, the combination of normalization and differences yielded more robust signals. Figure 3 illustrates that the combination of both pre-processing steps produces the most robust signal for predicting creatinine elevation by
amphotericin B, as measured by the number and magnitude of correctly oriented significant coefficients. In this, example both differences and normalization are required to identify a positive signal with confidence.

**Adding context-related variables to multivariate lagged linear regression**

We evaluated lagged drug coefficients for a multivariate autoregressive model that incorporates patients’ admission timelines as well as drug and lab measurement histories for all 20 drug-lab pairs using intra-patient normalized time series of differences. This allowed us to consider the extent to which contextual information, like inpatient admission, can recalibrate estimated effects of drugs on physiologic measurements. Using this method, we detected the same 6 known drug effects that were captured using the autoregressive drug and lab model. Little difference was seen in number or magnitude of significant drug coefficients between the autoregressive drug and lab model and the context model in cases of expected physiologic drug effects. However, the context model did produce significantly different results for some of the uncharacterized drug-lab pairs.

Figure 4a demonstrates that autoregressive context terms “explain away” the effect of simvastatin on creatinine, potassium, and hemoglobin (hemoglobin has only 1 significant coefficient in the context model) by adjusting the lagged drug coefficients to insignificant quantities, while keeping intact simvastatin’s propensity to increase Total CK via muscle damage. Figure 4b shows that admission does not, however, contribute additional information toward studying the effect of ibuprofen on the four considered lab tests. This result is striking, and suggests that inpatient admission is an important confounding variable to consider when analyzing temporal effects of simvastatin, but is largely unimportant for analyzing the effects of ibuprofen. While the significant relationships for ibuprofen may in fact be legitimate, it is equally possible that they are explained away by other process-related context variables.

**Discussion**

By developing a method for constructing time series of continuous and categorical variables, we were able to compare univariate and multivariate lagged regression models that incorporate lab measurements, drug orders, and inpatient admissions. All lagged methods showed highest specificity and sensitivity, overall, with intra-patient normalized laboratory values, and multivariate methods performed best in these metrics when differences were used during pre-processing stages. All multivariate methods identified the same six physiologic effects documented in clinical literature. Adding variables that describe patient context (lagged lab measurements and lagged admission events) increased the number and magnitude of significant drug coefficients in the expected cases and improved discrimination against unlikely
associations. We found that adding context-based variables to autoregressive models allowed for explicit handling of confounding variables and provided a simple way to evaluate the temporal effects of ordered drugs on physiology.

It is useful to note that the inpatient admission variables did not affect drug coefficients for warfarin or ibuprofen for most labs. This simply indicates that there is little correlation between admission, the drug, and each lab that cannot be explained by the drugs and labs alone. In this study, we are not interested in confounder coefficient magnitude or their relative explanatory contributions—rather, we are looking for the co-linearities that eliminate or enable rightfully significant drug coefficients (i.e. explaining away). Adding context-variables to autoregressive models appears to be a simple way to probe for confounders in EHR data without biasing the analysis. Selecting other possible confounders, like surgery, is a worthwhile exercise in better explaining the trends we observed with little physiologic interpretation. However, adding multiple new variables effectively adds new time points for interpolation in the time series, making each lag represent less average clock-time. It is unclear how many variables can be included in the timeline construction without detrimental distortion of results.

It is difficult to discuss the detected positive associations that are not reported in clinical literature. On one hand, data from the EHR are liable to bias, and confounders can only be accounted for explicitly in our formulation. On the other hand, subtle drug effects are likely to be unstudied, yet omnipresent in medical practice. For this reason, we do not formally compute specificity or sensitivity metrics. We were also unable to detect drops in hemoglobin to implicate ibuprofen and warfarin in bleeding events. Our results suggested that warfarin and ibuprofen increase hemoglobin levels, which contradicts strong clinical evidence that they are causative agents of intestinal bleeding as defined by low hemoglobin levels [23]. It may be that physicians are very careful to maintain normal or even high hemoglobin levels when concerned about bleeding. This is an open question that merits further investigation, and we hypothesize that introducing relevant context variables (e.g. surgery, other labs) in the autoregression could help untangle this problem. In general, different data sources may impose different limitations on inference. It is unclear why we did not observe a drop in hemoglobin, but it may be a natural limitation of the data we used, which may lack the measurement frequency necessary to capture short-term fluctuations in hemoglobin levels.

More broadly, we wish to better understand how temporal dependence between lagged variables (even after taking differences) manifests in the coefficient estimation. We found that changing the number of estimated lags often does not significantly alter the trend of coefficients. Specifically, similar trends for amphotericin B predicting elevated creatinine persist whether estimating 10, 30, or 60 lags. The plots, not shown, are all close to zero at their endpoints and peak in the middle. This is clearly an effect of co-linearity between lags of a particular variable, and it may be important to employ model selection methods, like Akaike information criterion [24] and Bayesian information criterion [25], to reduce the number of variables and their co-dependencies. These and other methods may assist in automating and optimizing feature selection (we hand-selected drugs and labs that are consistently taken across large patient cohorts at our center). Model selection criteria are also important for formally evaluating Granger causality between correlated variables, and we hope to have a method that rigorously evaluates causality.

We seek a reliable high-throughput method that produces meaningful and interpretable results that will allow us to uncover new associations—we can only build the methodology for detecting known associations de novo, but it is hard to know if such a methods will be as good at detecting previously hidden associations. Moreover, given that drug coefficients are lagged by sequence, it is difficult to map results back to actual clock-time. In principle, each lag $\tau$ is associated with a distribution of time lengths across all patients, and the summary statistics for those distributions may provide sufficient insight into the unique time-scale of each temporal process we detect.

Linear and non-linear distributed lag models have been used widely to study the temporal relationships between environmental factors and rates of health-related incidents, like suicide, mortality, and infectious disease [26,26–29], but their adoption in work with EHR data has been less common. These methods offer the advantage of explicit handling of confounding variables by including their autoregressive terms in the regression. More complex approaches also exist for removing confounders in distributed lag models, and Bahadori and Liu have proposed a methodology for applying Granger causality to medical data that is designed to learn the effects of unobserved confounders [30]. Ghassemi et al. used multi-task Gaussian process models for multivariate time series modeling of data collected in intensive care units (ICU) [31], and Joshi and Szolovits have applied novel unsupervised data mining techniques to characterize severity of patient physiology in ICUs [32]. Other noteworthy approaches for temporal pattern discovery have been applied to EHR data [5–7,33]. Jung and Shah evaluated the effect of non-stationarity in EHR data on different machine learning models, and found suboptimal performance of complex methods that ignore non-stationarity [34]. In addition, Lasko et al. have demonstrated machine learning methods that evaluate sampling rates and biases in laboratory measurements [35] and model temporal effects for both
continuous and categorical variables from the EHR [3,36]. While machine learning and pattern recognition methods can uncover complex relationships in high dimensional data, the lagged linear regression methods we use are advantageous due to their simplicity of implementation and interpretability.

This study was limited to EHR data from one medical center, and was limited to eight hypotheses based on review of clinical literature. However, there does not exist a ground truth for these hypotheses. Similarly, the twelve drug-lab pairs that were not found to be significantly linked in clinical literature cannot be verified and were thus exploratory, rather than formal negative controls.

### Conclusion

By comparing univariate and multivariate lagged regression models, we established methods for timeline construction that yielded consistent results across model implementations. We found that drug effects were best characterized, as compared to clinical literature, by multivariate lagged models that incorporate drug orders, laboratory measurements, and inpatient admission events for 20 example drug and lab pairs. These results suggest that simple autoregressive models of commonly available EHR data can be used to detect real physiologic drug effects in the presence of confounding health-care processes, and a more thorough study with a larger feature set is warranted.

### Table 1. Model performance*

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>lab,drug,context</th>
<th>lab,drug</th>
<th>lab,drug</th>
<th>lab,drug</th>
<th>lab,drug</th>
<th>lab,drug</th>
<th>lab,drug</th>
<th>Univariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>Lab</td>
<td>Expected</td>
<td>Citation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amphotericinB</td>
<td>Hemoglobin</td>
<td>0+,21-</td>
<td>0+,21-</td>
<td>0+,16-</td>
<td>1+,3-</td>
<td>0+,14-</td>
<td>0+,4-</td>
<td>0+,30-</td>
</tr>
<tr>
<td>amphotericinB</td>
<td>Total CK</td>
<td>0+,5-</td>
<td>0+,5-</td>
<td>0+,8-</td>
<td>1+,2-</td>
<td>0+,6-</td>
<td>0+,6-</td>
<td>0+,30-</td>
</tr>
<tr>
<td>amphotericinB</td>
<td>creatinine</td>
<td>pos [37]</td>
<td>22+,0-</td>
<td>23+,0-</td>
<td>21+,0-</td>
<td>1+,0</td>
<td>0+,0</td>
<td>19+,0</td>
</tr>
<tr>
<td>amphotericinB</td>
<td>potassium</td>
<td>neg [38]</td>
<td>0+,22-</td>
<td>0+,22-</td>
<td>0+,22-</td>
<td>1+,0</td>
<td>0+,20-</td>
<td>0+,10-</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>Hemoglobin</td>
<td>neg [23]</td>
<td>28+,0-</td>
<td>28+,0-</td>
<td>27+,0-</td>
<td>5+,0</td>
<td>0+,0</td>
<td>13+,0</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>Total CK</td>
<td>0+,8-</td>
<td>0+,8-</td>
<td>0+,1-</td>
<td>0+,2-</td>
<td>0+,0-</td>
<td>0+,1-</td>
<td>0+,4-</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>creatinine</td>
<td>pos [39,40]</td>
<td>26+,0-</td>
<td>26+,0-</td>
<td>30+,0-</td>
<td>2+,0</td>
<td>9+,0</td>
<td>18+,0</td>
</tr>
<tr>
<td>ibuprofen</td>
<td>potassium</td>
<td>Possible</td>
<td>pos [39,40]</td>
<td>23+,0-</td>
<td>24+,0-</td>
<td>23+,0-</td>
<td>7+,0</td>
<td>23+,0</td>
</tr>
<tr>
<td>simvastatin</td>
<td>Hemoglobin</td>
<td>[41]</td>
<td>1+,0</td>
<td>0+,2</td>
<td>0+,0</td>
<td>0+,3</td>
<td>0+,1</td>
<td>0+,22</td>
</tr>
<tr>
<td>simvastatin</td>
<td>Total CK</td>
<td>pos [42]</td>
<td>5+,0</td>
<td>8+,0</td>
<td>8+,0</td>
<td>2+,1</td>
<td>9+,0</td>
<td>7+,0</td>
</tr>
<tr>
<td>simvastatin</td>
<td>creatinine</td>
<td>0+,0</td>
<td>0+,4</td>
<td>0+,6</td>
<td>0+,0</td>
<td>0+,0</td>
<td>0+,5</td>
<td>0+,7</td>
</tr>
<tr>
<td>simvastatin</td>
<td>potassium</td>
<td>0+,0</td>
<td>0+,7</td>
<td>0+,11</td>
<td>1+,0</td>
<td>0+,0</td>
<td>0+,8</td>
<td>18+,0</td>
</tr>
<tr>
<td>spironolactone</td>
<td>Hemoglobin</td>
<td>28+,0-</td>
<td>27+,0-</td>
<td>28+,0-</td>
<td>2+,2</td>
<td>28+,0-</td>
<td>21+,0-</td>
<td>0+,5</td>
</tr>
<tr>
<td>spironolactone</td>
<td>Total CK</td>
<td>0+,0</td>
<td>0+,0</td>
<td>0+,2</td>
<td>0+,1</td>
<td>0+,0</td>
<td>0+,0</td>
<td>0+,0</td>
</tr>
<tr>
<td>spironolactone</td>
<td>creatinine</td>
<td>pos [43]</td>
<td>23+,0-</td>
<td>23+,0-</td>
<td>25+,0-</td>
<td>7+,0</td>
<td>2+,1</td>
<td>27+,0</td>
</tr>
<tr>
<td>spironolactone</td>
<td>potassium</td>
<td>pos [44]</td>
<td>28+,0-</td>
<td>28+,0-</td>
<td>29+,0-</td>
<td>8+,1</td>
<td>27+,0</td>
<td>27+,0</td>
</tr>
<tr>
<td>warfarin</td>
<td>Hemoglobin</td>
<td>neg [23]</td>
<td>30+,0-</td>
<td>30+,0-</td>
<td>30+,0-</td>
<td>10+,3</td>
<td>30+,0</td>
<td>29+,0</td>
</tr>
<tr>
<td>warfarin</td>
<td>Total CK</td>
<td>0+,4</td>
<td>0+,5</td>
<td>0+,16</td>
<td>0+,1</td>
<td>0+,0</td>
<td>14+,0</td>
<td>1+,12</td>
</tr>
<tr>
<td>warfarin</td>
<td>creatinine</td>
<td>28+,0-</td>
<td>27+,0-</td>
<td>29+,0-</td>
<td>7+,1</td>
<td>25+,1-</td>
<td>28+,0-</td>
<td>0+,18</td>
</tr>
<tr>
<td>warfarin</td>
<td>potassium</td>
<td>30+,0-</td>
<td>30+,0-</td>
<td>30+,0-</td>
<td>8+,2</td>
<td>28+,0-</td>
<td>18+,0</td>
<td>30+,0</td>
</tr>
</tbody>
</table>

*Pairs show the number of statistically significant positive and negative lags (e.g., “2+,1-” implies two positive lags and one negative lag). Green implies predominantly positive association, red implies predominantly negative, and grey implies minimal (less than 3) or mixed.

### Acknowledgment

This work was funded by National Library of Medicine grant R01 LM006910.

### References


A Topic-modeling Based Framework for Drug-drug Interaction Classification from Biomedical Text

*Dingcheng Li*¹, *Sijia Liu*¹², Majid Rastegar-Mojarak¹, Yanshan Wang¹
Vipin Chaudhary², Terry Therneau¹, Hongfang Liu¹
¹ Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
² Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA

Abstract

Classification of drug-drug interaction (DDI) from medical literatures is significant in preventing medication-related errors. Most of the existing machine learning approaches are based on supervised learning methods. However, the dynamic nature of drug knowledge, combined with the enormity and rapidly growing of the biomedical literatures make supervised DDI classification methods easily overfit the corpora and may not meet the needs of real-world applications. In this paper, we proposed a relation classification framework based on topic modeling (RelTM) augmented with distant supervision for the task of DDI from biomedical text. The uniqueness of RelTM lies in its two-level sampling from both DDI and drug entities. Through this design, RelTM take both relation features and drug mention features into considerations. An efficient inference algorithm for the model using Gibbs sampling is also proposed. Compared to the previous supervised models, our approach does not require human efforts such as annotation and labeling, which is its advantage in trending big data applications. Meanwhile, the distant supervision combination allows RelTM to incorporate rich existing knowledge resources provided by domain experts. The experimental results on the 2013 DDI challenge corpus reach 48% in F1 score, showing the effectiveness of RelTM.

Introduction

Potential drug-drug interactions, defined as the co-prescription of two or more drugs that are known to interact, are one of the primary causes of medical error¹². Some studies estimate that medical errors result in around 44,000 to 98,000 deaths per year in the United States³, and that 7,000 of those deaths are due to medication-related errors⁴. Therefore, it is significant to prevent or reduce such errors in order to improve the clinical diagnosis quality and thus save thousands of patients' lives. The Institute of Medicine has noted that a lack of drug knowledge is one of the most frequent proximal causes of such errors⁵. A report in Australia⁶ mentioned that 75% of hospital admissions related to medication errors are preventable, which implies that drug-drug interactions are a preventable cause of morbidity and mortality. Yet, the consequences of drug-drug interactions in the community are not well characterized⁶. Indeed, health care providers often have inadequate knowledge of what drug interactions can occur, of patient specific factors that can increase the risk of harm from an interaction, and of how to properly manage an interaction when patient exposure cannot be avoided⁷⁸.

One of the primary reasons for this knowledge gap is the lack of complete and authoritative source of DDI knowledge⁹. Rather, there are multiple sources, including DrugBank, DailyMed, National Drug File, Express Scripts DRUG DIGEST and Medscape for WebMD4, each tasked with extracting, evaluating, and staying up-to-date with pertinent DDIs reported in the literature, and drug product labeling¹⁰. However, the inconsistency of information found in these resources is a serious issue. A manual review by a clinical pharmacologist of 100 randomly selected potential interactions out of more than 300,000 automatically extracted interactions revealed that 40% were genuinely inconsistent¹¹. The dynamic nature of drug knowledge, combined with the enormity of the biomedical literature, makes the task of collecting and maintaining up-to-date information on drug-drug interactions extremely challenging and time-consuming.

There is a strong need to approach this task with automated methods, supplemented with human efforts. Natural language processing (NLP) and information extraction methods for identifying and extracting DDIs have been increased attention in the last few years, and several attempts have already been made to develop methods for this task, showing good potentials for success. In 2011, the first shared task challenge for DDI extraction,
DDIExtraction-2011 invited participants to develop automatic methods to extract DDIs. Most of participants developed binary classification models to handle this task. In the 2013 challenge, the organizers further classified the DDI categories and correspondingly, participants expanded their binary classification work into multiple classifications. Some of them took two steps to do the classification, first, detecting whether there is a relation among a pair of drugs and second, assign specific category to it. In general, those approaches yielded results ranging from 40% to 68%. However, all those approaches deployed supervised machine learning frameworks and focused more on feature engineering. Nonetheless, at the time when the data is growing exponentially, the supervised approaches cannot meet the needs of real-world applications. In particular, the dependency on annotated corpus forms a bottleneck for those methods since human annotations are expensive and time-consuming. Besides, the model will easily become over-fitted due to limited annotated corpus. On the other hand, if we only use unsupervised methods, like K-means, the results will be hard to interpret and therefore arduous to cluster into consistent categories.

In light of this, we conjecture that using semi-supervised or distant-supervised approach to detect and classify drug-drug interactions from biomedical text can be a better alternative because it can address limitations of both supervised and unsupervised methods. Specifically, we design, implement and evaluate a Bayesian model complemented with knowledge-driven distant supervision. In alignment with the term topic modeling for text mining, we call it relation topic modeling (RelTM). This approach attempts to make assumptions on the generative process in discourses and uses Gibbs sampling and Expectation-Maximization to infer the model parameters. Unlike supervised models such as support vector machines, or logistic regression, unsupervised Bayesian methods do not require human efforts such as annotation and labeling in data preparation stages, which is its advantage in today’s trending big data applications. Meanwhile, the distant supervision combination allows us to incorporate rich existing knowledge resources provided by domain experts.

Related Work

In this section, we discuss techniques related to both the DDI challenges and to our proposed framework. The state-of-art techniques for the DDI challenge participation in 2011 are from Segura et al. They used a hybrid method that combines supervised machine learning, shallow parsing and syntactic simplification with pattern matching. The UMLS MetaMap tool (MMT) was used to provide shallow syntactic parsing and a set of domain-specific lexical patterns were developed to extract DDIs. The experimental results showed that they achieved a precision of 0.51, a recall of 0.72 and an F1 score of 0.60 using SVM classifier. In the 2013 challenge, the system with the highest F1 score was proposed by the FBK-irst team. Their system is a multi-phase relation extraction system. They used two separate phases for DDI extraction and classification. For DDI extraction, they removed less informative sentences and instances, and then trained a system on the remaining instances. A hybrid kernel classifier that contains a feature based kernel, a shallow linguistic kernel, and a Path-Enclosed Tree kernel was used in the first step. For classification of DDI, they trained four separate models for each class (one vs. all the other classes). The innovative part of this system is detecting less informative sentences, where a sentence is considered less informative if all drugs in a sentence fall under the scope of a negation cue (such as not). A negation detector system (focused on a limited set of negation cues, such as no, n't and not) is used to identify and filter the less informative sentences. The remaining sentences are classified with the SVM Light-TK toolkit, utilizing the Charniak-Johnson reranking parser, a self-trained biomedical parsing model, and the Stanford parser. On the DDI-DrugBank test dataset, they obtained an F1 score of 0.68 and on the DDI-Medline test dataset, an F1 score of 0.40. Other high-performance systems usually took similar approaches with differences in the use of machine learning model and feature selections. All of them employed gold-standard labels for supervised learning besides the utilization of rich features and knowledge resources. Their model developments and especially their efforts in feature engineering provide us food for thought and allows us to directly explore those resources for our usage.

As mentioned in the introduction, in order to reduce the dependency on human annotations, we plan to develop semi-supervised models. Our proposed method is inspired by research in areas other than DDI classification. The work which is most relevant to us is the study of relation discovery from news domain using generative models. They proposed rel-LDA and type-LDA, the latter of which is quite similar to ours. Specifically, their type-LDA composed of a probability graphical model where each document comprises N pairs. Each pair consists of relation features, entity features for source argument and destination argument. This framework takes dependency path, trigger words, part-of-speech features and named entity tags as features. The pair here refers to a relation in the
document. Their relations include authorOf, bornIn, founder, parent and so on. Similarly, we have a generative graphical model to generate DDI, relation features and drug features.

Nonetheless, there are a few important differences in our work. First, our approach is a combination of generative model and clustering method. A similarity function is integrated into the relation sampling. This is inspired by Blei et al’s correlated topic models\textsuperscript{35} and distance dependent Chinese restaurant process\textsuperscript{36}, Ghosh et al’s spatial distance dependent Chinese restaurant process\textsuperscript{37} and Chang et al’s relation topic models for document networks\textsuperscript{38}. All of these techniques share the common point that they attempt to accommodate random partitions of non-exchangeable data by taking distances or similarities between topics into consideration. We notice that DDI pairs of same categories share correlated features, many of which are implicit. Consequently, methods based only on co-occurrence may miss clustering them together while the introduction of similarity functions may make up for this weakness. This is an innovative extension of type-LDA. With this extension, we can add rich knowledge resources and further we can flexibly experiment diverse similarity measures to discriminate relations more accurately.

Second, our work borrows the concept of distant supervision\textsuperscript{39}. Distant supervision for relation classification is based on the assumption that if two entities, which are relevant, are mentioned in one sentence, that sentence may imply a relation and thus can be used as evidence for that relation\textsuperscript{40}. Based on this assumption, we employ two knowledge sources, Drugbank and DailyMed, which are the two main resources of drug-drug interactions all over the world to boost the performance of our framework. Accordingly, our work shares some similarities with the work of Riedel et al\textsuperscript{11}, in which, they employ Freebase to train a semi-supervised graphical model to extract relations from New York Times corpus. Their semi-supervised graphical model is essentially an undirected factor-graph model, which makes use of Gibbs sampling to update confidence on a group of expressions involving two entities. With the confidence learning, a ranked list of expressions extracted from Freebase is used as evidence for relation extraction in New York Times corpus.

One more related work is our latest work on coreference resolution with non-parametric topic modeling. In that work\textsuperscript{42}, the proposed infinite mixture model integrates definite sampling with maximum likelihood estimation of mention similarity to estimate the cluster of coreferring entities in clinical notes. That system achieves an F measure of 0.847 on i2b2 2011 coreference dataset. Our current work shares two common ideas. First, both works attempt to handle relation classification (one for identity relation, the other for drug-drug interaction) under the framework of Bayesian topic modeling; second, similarity functions are utilized to guide the sampling of the objective functions. The difference lies in that our current work is not non-parametrical and distant supervision is employed. In addition, the generative process of our previous work is mention-entity sampling and the observed variable is the mention while our current work is feature-entity-relation sampling. More complex sampling is involved in our current approach.

Methods

The task of relation extraction can be abstracted as mapping surface textual relations to underlying semantic relations. There are several components to a coherent relation type, including a tight small number of textual expressions (two drug mentions, or called relation pair) as well as constraints on the entities involved in the relation. Often, the textual expressions themselves may not include sufficient information for us to model the relation type and thus make reasonable predictions. Hence, the constraints will play quite a big role in constructing discriminant predictors. For supervised approaches, those constraints are utilized as features to learn classifiers. Differently, in our work, we deploy them as observed variables for our Bayesian model plus the textual expressions for entities. Through this way, we can assume that multinomial distributions exist among those observed variables and further we can set Dirichlet priors for relation pairs. In addition, we notice that drug mentions, constraints and relation features belong to different semantic space where relation pairs depend on both relation features, such as trigger words and relation constraints, like negation tagger and drug mentions while drug mentions depend on constraints related to them, such as syntactic roles. Therefore, they should be modeled at different levels.

Based on above observations, we present relation topic modeling (RelTM), a semi-supervised probabilistic generative model for inducing clusters of relation types and recognizing their textual expressions. The set of relation types is not pre-specified but induced from observed unlabeled data. The relation induction will be augmented with distant supervision by incorporating external resources-based similarity functions.

We know that relations can only hold between certain entity types, known as selectional preferences (Ritter et al., 2010; Seaghdha, 2010; Kozareva and Hovy, 2010). This model can capture the selectional preferences of relations to their arguments (drugs in our context). In the mean time, it clusters pair into relational clusters, and arguments into
Instead of using the relations types from the labeled data, we suppose the data consists of more clusters, which is called latent relations in the following sections. In other words, we expect a specific relation type may have more than one pattern of feature set. This allows each relational cluster to have several sub-clusters. These sub-clusters (latent relations) are merged in the categorizing step.

Our observed data consist of a corpus of sentences and each sentence is represented by a bag of relation pairs. Each pair represents an observed drug-drug interaction between two drugs and consists of three components: the relation features, the source drug and the destination drug. The relation pair is the primary observed random variable in our model and we construct our models so that clusters consist of textual expressions representing the same underlying relation type. Since unlike classical LDA where observed words are independent and shared by the whole collection of documents, it will be reasonable to assume that each word may be generated by some specific topic. In our framework, a set of drug mention features supports one drug and relation features and mention features jointly support one relation. Therefore, the generative process is somewhat different. Namely, all of its features are only sampled under one specific relation and all of mention features are only sampled under one specific drug entity as well.

**Generative Process**

Before we start to introduce the generative process, some notations are explained. The symbol $\sim$ stands for drawing the sample given a specific distribution. For example, the notation $x \sim \text{Multi}(\theta)$ represents sampling variable $x$ from

<table>
<thead>
<tr>
<th>Notation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>The total number of sentences</td>
</tr>
<tr>
<td>$</td>
<td>R</td>
</tr>
<tr>
<td>$N$</td>
<td>A random variable to represent the total number of relation pairs in a sentence</td>
</tr>
<tr>
<td>$</td>
<td>F</td>
</tr>
<tr>
<td>$\alpha_{1,R}$</td>
<td>R-dimensional parameter of a Dirichlet distribution for relations</td>
</tr>
<tr>
<td>$\beta_{1,R}$</td>
<td>Parameters for R component distribution over relation features</td>
</tr>
<tr>
<td>$\beta_{1,R}^{[w,v]}$</td>
<td>Parameters for R component distribution over drug entity 1 ($w$) or entity 2 ($v$)</td>
</tr>
<tr>
<td>$\beta_{1,F}^{[y,z]}$</td>
<td>Parameters for F component distribution over features of drug entity 1 ($y$) or entity 2 ($z$)</td>
</tr>
<tr>
<td>$\theta_{1,R}$</td>
<td>R-dimensional parameters of relation distribution variables over a sentence</td>
</tr>
<tr>
<td>$\phi_{1,R}$</td>
<td>Parameters for R component distribution over the relation types</td>
</tr>
<tr>
<td>$\phi_{1,R}^{[w,v]}$</td>
<td>Parameters for R component distribution over drug entity 1 ($w$) or entity 2 ($v$)</td>
</tr>
<tr>
<td>$\phi_{1,F}^{[y,z]}$</td>
<td>Parameters for F component distribution over features of drug entity 1</td>
</tr>
<tr>
<td>$r_{1,N}$</td>
<td>Variables representing a sequence of relations in a specific sentence</td>
</tr>
<tr>
<td>$f$</td>
<td>Features of a given relation pair</td>
</tr>
<tr>
<td>${w, v}_{1,N}$</td>
<td>Variables representing hidden variable, the drug entity 1($w$) or entity 2($v$) in a specific sentence</td>
</tr>
<tr>
<td>${y, z}_{1,R}$</td>
<td>Finite set of observed variables that represents a specific features of drug entity 1($w$) or entity 2($v$)</td>
</tr>
</tbody>
</table>

**Figure 1 The Plate Diagram of DDI-LDA**
the multinomial distribution given parameter \( \theta \).

Correspondingly, 
\( x \sim \text{Dirichlet}(\alpha) \) stands for sampling variable \( x \) from the multinomial distribution given parameter \( \alpha \). Table 1 shows other notations used in this paper. Following the convention of NLP community, the real world objects mentioned in a discourse by noun phrases or pronouns are usually called entities, in our context, drugs, and the phrases or pronouns appearing in discourses are called mentions, namely drug instances. Mentions may consist of several words in free texts and each of such words is called a token. Each relation is composed of two drug mentions.

The plate diagram of graphical model of RelTM is illustrated in Figure 1. The generative process of relation features, drug entities and their entity features are described in Figure 2. Formally, there are \( N \) numbers of possible relations \( (N \) is a random variable) in each sentence. Each relation is represented by three distributions, namely, relation feature distribution, features of drug entity 1 and features of drug entity 2. When we detect the relations between two drugs, we first choose a relation type based on a relation type distribution, composed of \( |R| \) number of distribution with summation 1. Then we generate a bag of relation features one by one based on the chosen latent relation type as well as generate a pair of drug entities. Sequentially, drug entity 1 and the drug entity 2 are employed to generate mention features for specific drugs, namely, instances of entities. Essentially, we do all sampling with Dirichlet for priors and multinomial for their conjugate posteriors. Namely, each multinomial distribution is governed by symmetric Dirichlet distribution.

The features of a relation pair consist of relation level features and mention level features. Relation level features include part-of-speech, dependency path, trigger, drug-bank indication (indications of drug-drug relations from the drug-bank database\(^{13}\)), lexicon and POS features, mention level features include the drug mention itself, its named entity tag, \( i.e. \) the type of drug and its Unified Medical Language System (UMLS) concept unique identifier (CUI) and type unique identifier (TUI).

**Strengthened Model with Similarity Measures**

Although the current framework takes both relation features and mention features into considerations, it is found that co-occurrences hypothesis, the foundation of topic modeling, may lead to poor performance while evaluated by DDI evaluation metrics. Therefore, similarity measures are introduced to improve the system performance. We use Simple Matching Coefficient (SMC)\(^2\) as the similarity measure of two different relation pairs, which is widely used in data mining and statistics for quantifying the similarity of two binary vectors. The SMC of two different pairs can be simply represented by:

\[
SMC(w, v) = \frac{\text{Number of matching features of } w \text{ and } v}{\text{Number of total features of } w \text{ and } v}
\]

Equation 1

Specifically, after each sampling of relations for one pair of drugs is done, we selected high-discriminate features to compute similarity scores between current relation pair and previous relation pairs, instead of between two large feature sets. In this study, the trigger words in Drugbank and DailyMed are used. In other words, if two drug entities

---

1. For each sentence, draw a component proportion \( \theta^s \sim \text{Dirichlet}(\alpha) \).
2. For each pair \( p = 1, \ldots, N_p \), draw \( r^p \sim \text{Multi}(\theta^p) \)
   a. For each relation feature \( n = 1, \ldots, N_n^p \)
      i. Draw \( \varnothing_f \sim \text{Dirichlet}(\beta_f) \)
      ii. Draw \( f_p,n \sim \text{Multi}(\varnothing_f) \)
   b. For entity \( t = 1, \ldots, N_e \)
      i. Draw \( \varnothing_w \sim \text{Dirichlet}(\beta_w) \)
      ii. Draw \( w_t \sim \text{Multi}(\varnothing_w, r) \)
      iii. For each entity feature \( n = 1, \ldots, N_n^w \)
          1. Draw \( \varnothing_y \sim \text{Dirichlet}(\beta_y) \)
          2. Draw \( y_{w,t} \sim \text{Multi}(\varnothing_y) \)
   c. For entity \( i = 1, \ldots, N_e \)
      i. Draw \( \varnothing_v \sim \text{Dirichlet}(\beta_v) \)
      ii. Draw \( v_t \sim \text{Multi}(\varnothing_v, r) \)
      iii. For each entity feature \( n = 1, \ldots, N_n^v \)
          1. Draw \( \varnothing_z \sim \text{Dirichlet}(\beta_z) \)
          2. Draw \( z_{v,t} \sim \text{Multi}(\varnothing_z) \)

**Figure 2** The generative process of DDI-LDA
are mentioned with the same trigger verb in these two corpora, the similar feature count is incremented by one. After the count for each pair is obtained, it is divided by the size of the union of the entity feature sets. This step performs normalization to the co-occurrence feature count. Then the normalized feature counts of given relation assumptions are added to the probability vector in the inference stage.

Inferences

For simplicity and clarity, we omit the likelihood functions and their detailed derivations in this paper. We use collapsed Gibbs sampling to perform model inference. In collapsed Gibbs sampling, the distribution of a relation for its related features and two drug mentions are based on values of mention features is computed. For the simplicity of the equation, we first define the following function:

$$\Gamma(x, K) = \frac{n^i_{t-i} + \beta^i_t}{\sum_{i=1}^{\|F\|}(n^i_{t-i} + \beta^i_t)}$$  \hspace{1cm} \text{Eq. 2}

where $x = \{y, z, w, v, f\}$ and $K = \{|F|, |F|, |R|, |R|, |R|\}$ respectively. Then we have:

$$pp(r_t|r_{-t}, x, \alpha, \beta^x) \propto \frac{n^p_{m,i} + \alpha_m}{\sum_{m=1}^{\|P\|}(n^p_{m,i} + \alpha_m)} \prod_{x=(y,z,w,v,f)} \Gamma(x, K) \cdot \text{SMC}(w, v)$$  \hspace{1cm} \text{Eq. 3}

In this Eq., $n^p_{m,i}$ is a count of how many features are assigned to relations in sentence $s$, excluding the count of current relation $i$. The first term is the distribution of current relation $i$ sampled from the prior $\alpha$. The second term is the product of distributions of a series of supporting feature vectors. They include the proportion of the mention features for drug entity 1 and for drug entity 2, the proportion of the drug entity 1 and drug entity 2 and the proportion of relation features in current relation $i$ sampled from the prior $\alpha$. And the third term is SMC term. For the simplicity of the derivations of Gibbs sampling, we separate the RelTM term and the SMC term in inference stage. That is to say, the RelTM is optimized during the inference stage, while the SMC term remains constant in Gibbs sampling.

$$p(w_t = j|w_{-t}, r, \tilde{y}) \propto \frac{n^i_{t-i} + \beta_{t1}}{\sum_{i=1}^{\|W\|}(n^i_{t-i} + \beta_{t1})} \frac{n^j_{t-i} + \beta_{w_t}}{\sum_{i=1}^{\|W\|}(n^j_{t-i} + \beta_{w_t})}$$  \hspace{1cm} \text{Eq. 4}

$$p(v_t = j|v_{-t}, r, z) \propto \frac{n^i_{t-i} + \beta_{t2}}{\sum_{i=1}^{\|W\|}(n^i_{t-i} + \beta_{t2})} \frac{n^j_{t-i} + \beta_{v_t}}{\sum_{i=1}^{\|W\|}(n^j_{t-i} + \beta_{v_t})}$$  \hspace{1cm} \text{Eq. 5}

The above two equations compute the distribution of the two drug entities. Similar to Eq. 2, the distribution of drug entity 1 is based on the counts of features assigned to all entities in sentence $s$, excluding the count of current entity $i$ (term 1) and based on the distribution of all drug entities in sentence $s$ (term 2). Similar is the distribution of entity 2.

Categorizing topics

For the DDI Drugbank dataset, because the ground truth relations have been manually labeled, intuitively, if an inferred latent relation is associated with many pairs in a particular category, the relation is likely to belong to that category. To capture this intuition, we match the inferred latent relations to category $q$ where $q = arg\max_p p(q|r) = arg\max_q p(r|q) * \frac{p(q)}{p(r)} = arg\max_q p(r|q)$, assuming that all categories are equally important. We can estimate the probability of inferred latent relation type given category $q$ as,

$$p(r|q) = \frac{\sum_{s \in S_q} p(r|s)}{|S_q|}$$  \hspace{1cm} \text{Eq. 6}

where $p(r|s)$ denotes the learned probability of relation $r$ given sentence $s$ and $S_q$ denotes the subset of sentences in the DDI drugbank data collection that are labeled as category $q$.  

794
Experimental Results

As a gold standard for all experiments in this task, we used DDIExtraction-2013 corpus provided by the challenge organizer, for development. It contains 142 Medline abstracts on the subject of drug-drug interactions, and 572 documents describing drug-drug interactions from the DrugBank database. The corpus includes 6976 sentences that were annotated with four types of pharmacological entities and four types of DDIs. The DDIs types are: advice, effect, mechanism, and int.

The four types of drug-drug-interactions are defined as follows.

- Advice: the sentence notes a recommendation or advice related to the concomitant use of the two drugs. For example, "... UROXATRAL should NOT be used in combination with other alpha-blockers."
- Effect: the sentence states the effect of the drug interactions, including pharmacodynamic effect or mechanism of interaction. For example, "Quinones may enhance the effects of the oral anticoagulant, warfarin, ..."
- Mechanism: the sentence describes a pharmacokinetic mechanism.
- Int: the sentence mentions a drug interaction but doesn't provide any additional information.

To demonstrate the effectiveness of our proposed model, we conducted a series of experiments on DDI Drugbank dataset. In our problem setting, we assume that all drugs are provided and sentences are the place all interaction among drugs take place. Any two drugs among sentences are potential relation pairs. Yet, we do not suppose relation labels between any two drugs in a sentence of the training data are known. Namely, we target to develop semi-supervised framework to assign labels to those unlabeled DDI pairs with the four categories introduced in Table 2.

Since we are using an unsupervised method in model development, we did not divide our dataset into training set and testing set. That is to say the experiments are carried out in the corpus described in Table 2. There are two models evaluated in this section: baseline RelTM model and the RelTM with similarity measure (SMC) model. The performance metrics is computed with the package provided by DDIExtraction 2013 Challenge organizers. The performance of the two models mentioned above is shown in columns 3 to columns 5 and columns 6 to columns 8, respectively. Comparing these results, we can draw the conclusion that the combination of RelTM and distant supervision together improves the overall performance, which is indicated in the micro-average F1 score, from 0.3624 increasing to 0.480.

### Table 2 Performance evaluation in DDI Drugbank dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Number of pairs</th>
<th>RelTM baseline</th>
<th>RelTM+SMC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Mechanism</td>
<td>827</td>
<td>0.325</td>
<td>0.302</td>
</tr>
<tr>
<td>Effect</td>
<td>1700</td>
<td>0.495</td>
<td>0.536</td>
</tr>
<tr>
<td>Advise</td>
<td>1322</td>
<td>0.386</td>
<td>0.210</td>
</tr>
<tr>
<td>Int</td>
<td>188</td>
<td>0.135</td>
<td>0.526</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.355</td>
<td>0.393</td>
<td>0.362</td>
</tr>
</tbody>
</table>

Since we are using a model with pre-specified number of latent relations, the same as other finite mixture model like LDA, the number of latent relations (topics) will also influence the system performance. Figure 3 demonstrates the relation between number of specified latent relations of our model and the micro-average F1 score. From the curve there are three trends worth mentioning. The first trend is when the number of latent relations is smaller than 15, the F1 score will increase with the increase of number of latent relations. This is because when the number of latent relations is less than the optimal number, the model may fail in splitting some similarity patterns of different DDI pairs. The optimal number of latent relations observed in the dataset is 15, based on experiments. The second trend is, after the optimal latent relation number, the F1 score will decrease with the number of relations increasing. This is due to the decreasing effectiveness of Gibbs sampling method we are using for the model inference. A similar pattern can also be found in conventional LDA model. However, as we continually obtained finer and finer clusters of DDI pairs, which applies after the relation number is larger than 50, the F1 score will slightly increase. This trend...
is because of the greedy manner of how we do the cluster labeling. As the clusters become smaller, the accuracy of cluster labeling will benefit the F1 score.

**Discussion**

Due to the characteristic of Bayesian inferences, all the clusters are generated as random variables. It means they are generated according to the predicted probabilities of random variables in each category, instead of being directly assigned a label as in supervised methods such as support vector machine or random forest. The output results show our system can detect most of the strong category patterns, like the existence of the word “should” in *advise* and “effective” in *effect* in given sentences. However, while sampling the category label using category probabilities, most of the elements in the probability vectors are not strictly zero. Consequently, there is still possibility that DDI pairs with these strong patterns are not correctly classified, but the possibility of this kind of errors is within a low and acceptable range.

Another issue impacting the overall performance is the unbalanced category, in our study, *int*. In Table 2 we can find that there are only 188 instances of *int* pairs, which is much smaller than the second smallest category, which has 827 instances. Without any prior knowledge of category proportion, it is difficult for an unsupervised model to generate this small cluster, compared with other tunable supervised methods. As a result, our system generated more false positives in *int* categories than the other relatively large clusters. Therefore, the precision is lower that other categories, and it further impairs the micro-average F1 score. If we can utilize the prior knowledge of expected category proportion, this can be avoided by using unbalanced vector $\alpha$ as Bayesian prior of RelTM instead of a scalar value we are currently using.

Although the performance of our system is not as good as the state-of-art supervised approaches, it does induce promising results, which shows that the utilization of distant supervised machine learning methods like LDA is promising in relation classification, even in challenging tasks as DDI classification. With the augmented similarity measures, which use domain knowledge in medical corpora, the results are improved. It is reasonable to believe that if more prior knowledge is used in the proposed model, then it has a potential to detect relations more accurately. This paves the way for us to deploy and extend our current system on large-scale unlabeled medical literatures and thus reduce the dependencies on the corpus annotations.

**Conclusion and Future Work**

In this paper, we proposed a topic modeling based distant supervised approach for the task of drug-drug interaction from biomedical text. Specifically, we proposed to design, implement and evaluate a Bayesian model complemented with knowledge-driven distant supervision. Under the assumption that the relations of drug-drug interaction are generated from proposed generative process, an inference algorithm using Gibbs sampling was also designed. A similarity measure was also introduced to strengthen the model and its effectiveness is proven by the experiments. The proposed approach does not require human efforts such as annotation and labeling in data preparation stage, which is its advantage in today’s trending big data applications. Meanwhile, the distant supervision combination allows us to incorporate rich existing knowledge resources provided by domain experts.

In future, we will address problems found in error analysis. In addition, we will extend the similarity function by incorporating updating rules and correlation between latent relations. One direction of exploration includes the release of symmetric Dirichlet prior to non-symmetric considering the unbalanced numbers of relation categories.
Acknowledgements

The authors gratefully acknowledge the support from the National Institute of Health (NIH) grant 1R01LM011934 and 1K99LM012021-01A1.

Reference

29. Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. EACL; 2006.
Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation

Xiang Li, PhD¹, Haifeng Liu, PhD¹, Xin Du, MD², Ping Zhang, PhD³, Gang Hu¹, Guotong Xie¹, Shijing Guo¹, Meilin Xu⁴, Xiaoping Xie⁴

¹IBM Research - China, Beijing, China
²Department of Cardiology, Beijing Anzhen Hospital, Beijing, China
³IBM T.J. Watson Research Center, New York, USA
⁴Pfizer Investment Co. Ltd., Beijing, China

Abstract

Atrial fibrillation (AF) is a common cardiac rhythm disorder, which increases the risk of ischemic stroke and other thromboembolism (TE). Accurate prediction of TE is highly valuable for early intervention to AF patients. However, the prediction performance of previous TE risk models for AF is not satisfactory. In this study, we used integrated machine learning and data mining approaches to build 2-year TE prediction models for AF from Chinese Atrial Fibrillation Registry data. We first performed data cleansing and imputation on the raw data to generate available dataset. Then a series of feature construction and selection methods were used to identify predictive risk factors, based on which supervised learning methods were applied to build the prediction models. The experimental results show that our approach can achieve higher prediction performance (AUC: 0.71–0.74) than previous TE prediction models for AF (AUC: 0.66–0.69), and identify new potential risk factors as well.

Introduction

Atrial fibrillation (AF) is one of the most common clinical arrhythmias, affecting approximately 4 million adults in China¹. AF significantly increases the risk of ischemic stroke and other thromboembolism (TE). Moreover, compared to non-AF ischemic stroke, AF related ischemic stroke is more fatal and disabling². Oral anticoagulation (OAC) including warfarin has shown great efficacy in preventing ischemic stroke and TE for AF patients³. However, because OAC may have severe side effects such as warfarin bleeding, it is normally only recommended to AF patients with high risk of TE in clinical guidelines⁴. Besides, though radiofrequency ablation (RFA) is an effective procedure to treat AF and then reduce the risk of TE, it is still a scarce medical resource in present day China and increases economic burden on AF patients. Therefore, it is critical to accurately predict the risks of TE for AF patients and identify those truly high risk patients that should be treated by OAC and/or RFA.

Current ischemic stroke and TE risk models for AF, such as CHADS²⁶, CHA²DS²-VASc⁷ and Framingham Score⁸, were developed to stratify AF patients into categories of high, intermediate, and low risk. The risk factors used in these models, such as age, gender, prior ischemic stroke and TE, hypertension, diabetes, congestive heart failure (CHF), etc., are grounded in previous known evidence and experience, which are well understood and easy to apply. However, these risk models have only moderate prediction performance⁹ (the area under the receiver operating characteristic curve (AUC) is usually less than 0.7¹). It is mainly because that some potential risk factors that are highly related to TE occurrence for AF patients were not previously identified and involved in these risk models.

The objective of this study was to build 2-year ischemic stroke and TE prediction models for AF with high prediction ability and interpretability, based on the Chinese Atrial Fibrillation Registry (CAFR) data. The CAFR study started from the year of 2011, and has enrolled more than 17,000 AF patients from 32 hospitals in Beijing, China. The study collected the patients’ demographics, symptoms and signs, medical history, results of physical examination and laboratory test, details of treatments at baseline, and followed up the patients every 6 months. At every follow-up visit, the clinical events such as ischemic stroke and TE were collected.

Many previous works used statistical inference and machine learning methods to build high accuracy risk prediction models for patients with cardiovascular, diabetes and other diseases¹⁰,¹¹,¹²,¹³. However, it is still a challenging problem to build accurate and clinically interpretable TE prediction models from CAFR data. The first challenge is from the dataset, which is heterogeneous, non-standardized, incomplete and redundant. Much elaborate data curation work has to be done to remedy the data before analysis. Also, feature engineering should be performed to transform data types...
and reduce the redundancy of features. Another challenge is from the requirement of interpretability. Since we wanted to build human understandable and applicable prediction models, the dimensionality reduction and learning algorithms in which resulting models are difficult to interpret (e.g., principle component analysis and support vector machine) were not preferable.

In this paper, we address these issues by using integrated machine learning and data mining approaches to build TE prediction models for AF patients. We first performed data cleansing and imputation on the raw CAFR dataset to standardize the features and fill-in missing entries. Then a series of feature construction and selection methods were used to identify predictive risk factors and reduce redundancy. Finally, we applied different categories of supervised learning methods that have good interpretability, including generalized linear model, Bayes model and decision tree model, to build TE prediction models for AF. The experimental results show that our approach can achieve higher prediction performance than previous TE risk models, and also identify new potential risk factors that have not been previously identified or commonly used.

Methods

Figure 1 shows our approach pipeline of building ischemic stroke and TE prediction models for AF patients from CAFR data. We first selected and constructed the patient cohort of interest. Then data curation, including data cleansing and missing data imputation, were performed to generate available dataset. After that, we applied a series of feature engineering methods, including feature construction and feature selection, to identify the potential risk factors for predicting TE in AF. Finally, we trained prediction models using different supervised learning algorithms, and evaluated their prediction performance in terms of AUC and the area under the precision recall curve (AUPR) by cross validation and train/test splitting.

**Cohort**

The purpose of TE prediction models is to help clinicians identify AF patients with high risk of TE, and then decide to whom relevant interventions such as OAC and RFA should be used. Therefore, in this study the patients of interest are those who had not been treated with OAC (mainly warfarin in our data) or RFA at baseline. From the CAFR data, we identified 1864 AF patients who meet this criteria, where 193 patients (10.4%) are cases who had TE within 2 years after baseline, and 1671 patients are control instances who completed 2-year follow-ups and did not have TE within 2 years. The features used in our analysis include demographics, symptoms and signs, medical history, vital signs, laboratory test results, life styles and treatments.
Data Curation

In the raw CAFR data, more than half of the features have non-standardized and dirty values, and this percentage for the numeric type is particularly higher. Besides, the raw data has significant omissions due to the questionnaire structure, unknown values or errors in data collection. The workload of manually correcting these dirty and missing values could be enormous. Therefore, we performed automatic data curation, including data cleansing and missing data imputation, before predictive analysis.

1) Data Cleansing

The raw data is heterogeneous and includes different types of data items: binary type (e.g., hypertension history), nominal type (e.g., AF type) and numeric type (e.g., systolic blood pressure, SBP). For different data types, we designed sets of cleansing rules to remedy the non-standardized and dirty values in batch. These cleansing rules can be used to standardize data formats, correct input errors, or discard the values that cannot be recognized as the target types.

For the numeric features, some values in the raw data are not standardized (e.g. the full-width Chinese character “。” in Figure 2(a.1)), so we first defined a set of cleansing rules to standardize these values (e.g., Figure 2(b.i) that transforms “。” to the half-width character “.”). Besides, because the dataset was collected from different hospitals, a numeric feature may have different units (e.g., mg/dL and µmol/L for serum creatinine in Figure 2(a.2-3)). Therefore, we defined cleansing rules to unify these units (e.g., Figure 2(b.ii) that transforms mg/dL to µmol/L for serum creatinine). Finally, we discarded the numeric values that are out-of-range (e.g., the SBP greater than 200 mmHg in Figure 2(a.4)) and the non-numeric values (e.g., the values in Figure 2(a.5-6)) in numeric columns. Figure 2(c) shows the cleansed numeric values of the examples in Figure 2(a). For the binary features and nominal features, we also defined corresponding cleansing rules to standardize the formats and discard the unstructured values.

2) Missing Data Imputation

Data imputation is the process to remedy missing data, which is usually necessary for building a reasonable prediction model. Since the dataset is derived from questionnaire and the features have interrelationships, we first built a set of imputation rules to infer the missing values from the other relevant features. For example, if a patient’s SBP was greater than 140 mm Hg, then his/her hypertension history can be replaced with “true”. For another example, if the value of a parent question “having heart failure symptoms” is “false”, then all its children questions, such as “cantering rhythm”, can be imputed with “false”.

After that, we statistically imputed the remaining missing values that cannot be inferred from other features. We first discarded the features with too many missing entries, because their distributions are difficult to estimate, which may lead to inaccurate imputation results. Concretely, if a binary feature has more than 80% missing instances, or a numeric/ multi-value nominal feature has more than 60% missing entries, then this feature was removed from the dataset. For the remaining features, every missing value of a numeric feature (e.g., SBP) was replaced with the mean of the feature’s observed values, every missing value of an ordinal feature (e.g., NYHA level) was replaced with the median of its observed values, and every missing value of an unordered nominal feature (e.g., AF type) was replaced with the mode of its observed values. These methods can statistically minimize the impact of the imputed values in predictive modeling.
Risk Factor Identification

Risk factors for ischemic stroke and TE in AF have been previously studied\(^{6,7,8}\). These knowledge-based risk factors are grounded in previous evidence and have good interpretability. However, many other risk factors that are highly related to TE occurrence for AF patients were not previously identified and involved, and the models built from the previously known factors may not have adequate predictive power.

On the other hand, feature selection methods in machine learning\(^14\) can be used to automatically test and select predictive features from a large number of candidate features, and discover new risk factors that have not been previously identified. The prediction models built from the automatically identified factors can represent more complex disease progressions, and usually have higher predictive performance than the models derived from the knowledge-based factors\(^{10,12}\). The disadvantage of the data driven methods is that the resulting models may usually be difficult to interpret or apply in real clinical practices, because the original features in the data are not as easy to understand as the knowledge-based factors. To address these problems, we first performed feature construction to transform the original features and combine knowledge-based features. Then feature selection algorithms were applied to identify potential risk factors from the original and knowledge-based features.

1) Feature Construction
We first preformed feature transformation to split each multi-value feature to a set of binary features. After that, a set of formulas provided by clinicians were used to generate knowledge-based combination features, which can also be used as candidate features in feature selection. These knowledge-based features describe high-level clinical concepts, and each of them maps to multiple original features in the CAFR data. For example, in this study, “CHF” is defined by four features: “NYHA level” > 2 or “left ventricular ejection fractions” < 40% or “having heart failure history” or “having heart failure symptoms”. Similarly, “diabetes mellitus” is defined as “glycated hemoglobin (HbA1c)” ≥ 6.5% or “fasting plasma glucose” ≥ 7.0 mmol/L or “having diabetes history”.

2) Feature Selection
Before feature selection, we first performed pre-selection to remove the unreasonable features. We asked the clinicians to select the feature categories of interest, and discarded the uninteresting features (e.g., all subjective features about quality of life were discarded in order to avoid bias). The close-to-constant features, in which 99% of the instances have identical values, were also be removed.

In machine learning, there are three main supervised feature selection strategies: filter, wrapper and embedded optimization\(^{14}\). In this study, we employed and compared these methods in identifying predictive risk factors.

- **Filter.** This category of methods calculates a score to represent the relevancy of a feature (or a group of features) against the outcome, and then filters the features based on the score. In this study, we applied two univariate filter methods, which respectively use the p-value from chi-squared test and the information gain as the relevancy score to filter each feature independently. We also used the correlation-based feature subset selection method\(^{15}\) (CFS), which is a multivariate filter method that evaluates features in a batch way, to obtain the subset of features that are highly correlated with the outcome while having low intercorrelation between the features.

- **Wrapper.** This type of methods utilizes a specific classifier (e.g., logistic regression) to select the subset of features that provides the best performance for a specific metric (e.g., AUC). In this study, we applied the wrapper subset selection method\(^{16}\). This method evaluates a subset of features by the prediction performance of the classifier using cross validation, and uses the best first search strategy to search the subset of features that can achieve optimized performance.

- **Embedded optimization.** These methods incorporate feature selection directly into the learning process of a model. In this study, we used Lasso\(^{17}\) to introduce L1-norm regularization to generalized linear models (e.g., logistic regression), which can achieve feature selection by shrinking the coefficients of low relevant features to zero during model training.

Predictive Modeling
In this study, we applied and compared different categories of machine learning models that have good interpretability, including generalized linear models, Bayes models and decision tree models, to build TE prediction models for AF.
• **Generalized linear model (GLM).** GLM generalizes ordinary linear regression by allowing the linear model to be related to the response variable via a link function. We used logistic regression, which is a GLM with a logit link function and a binomial distribution. It was widely used in both medical statistics and machine learning due to its good performance and interpretability. We also applied Cox proportional hazards model\textsuperscript{18} in this study, which is a statistical model commonly used in survival analysis. Cox is a semi-parametric GLM that takes into account the time of observations.

• **Bayes model.** Naive Bayes model\textsuperscript{19} is a probabilistic classifier based on Bayes theorem with strong independence assumptions between the features. Naive Bayes models can be trained very efficiently, and have decent prediction performance and interpretability as well.

• **Decision tree model.** The classification and regression tree (CART) method was applied to build tree-based prediction model, which is very easy to interpret. We also employed random forest\textsuperscript{20}, which constructs a multitude of decision trees and outputs the mode of the classes of the individual trees. Compared to other decision tree learning methods, random forest generally has greater performance by reducing the problem of over-fitting, though having worse interpretability.

**Results**

We evaluated the performance of our approaches, including data curation, feature engineering and supervised learning, in building 2-year TE prediction models for AF patients from CAFR dataset. Figure 3 demonstrates the proportions of dirty and missing data in our dataset before and after data curation. In the 221 original features of the raw data, 116 features (55.0%) have non-standardized and dirty values, and 211 features (95.5%) have missing values (Figure 3(a)). After data cleansing, the dirty data in 91 features were standardized and/or corrected (Figure 3(b)). And after data imputation, the missing data in 155 features were filled-in, while 54 features with dirty values and/or missing values that cannot be remedied were discarded (Figure 3(c)). Finally, 167 available features were produced to the following feature engineering step.

Two standard metrics, AUC and AUPR, were used to evaluate the prediction performance of models. We used AUPR in addition to AUC because in our case of imbalanced dataset, AUC may provide an overly optimistic view of performance, while AUPR can provide a more informative assessment under this situation\textsuperscript{21}. Note that the baseline of AUPR for our dataset is 0.104, which is the average precision of randomly predicting the risk (i.e., an AUPR of 0.208 means the average precision is doubled than that of random prediction).

![Figure 3. Statistics of data quality before and after data curation](image-url)
Table 1. Mean and standard deviation of AUC and AUPR of the logistic regression models built on different feature sets, evaluated by cross validation. The standard deviation of every AUC and AUPR is less than 0.01.

<table>
<thead>
<tr>
<th>Candidate Features:</th>
<th>Original features</th>
<th>Original + knowledge-based features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection Method</td>
<td>No.</td>
<td>AUC</td>
</tr>
<tr>
<td>None</td>
<td>107</td>
<td>0.649</td>
</tr>
<tr>
<td>Chi-squared filter (p &lt; 0.001)</td>
<td>21</td>
<td>0.711</td>
</tr>
<tr>
<td>Information gain (IG &gt; 0.001)</td>
<td>36</td>
<td>0.697</td>
</tr>
<tr>
<td>CFS</td>
<td>16</td>
<td>0.722</td>
</tr>
<tr>
<td>Wrapper for AUC</td>
<td>23</td>
<td>0.759</td>
</tr>
<tr>
<td>Lasso (C = 0.1)</td>
<td>22</td>
<td>0.716</td>
</tr>
</tbody>
</table>

Table 2. Average AUC and AUPR of different learning models on different feature sets, evaluated by cross validation. The standard deviation of every AUC and AUPR is less than 0.02.

<table>
<thead>
<tr>
<th>Selection:</th>
<th>None</th>
<th>Chi-squared filter</th>
<th>CFS</th>
<th>Wrapper for AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Models</td>
<td>AUC</td>
<td>AUPR</td>
<td>AUC</td>
<td>AUPR</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.645</td>
<td>0.177</td>
<td>0.709</td>
<td>0.209</td>
</tr>
<tr>
<td>Cox</td>
<td>0.621</td>
<td>0.173</td>
<td>0.707</td>
<td>0.208</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.689</td>
<td>0.194</td>
<td>0.706</td>
<td>0.226</td>
</tr>
<tr>
<td>CART</td>
<td>0.653</td>
<td>0.171</td>
<td>0.645</td>
<td>0.173</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.696</td>
<td>0.203</td>
<td>0.708</td>
<td>0.210</td>
</tr>
</tbody>
</table>

To compare the performance of the feature engineering methods, we built logistic regression models on different feature sets generated by different feature construction and feature selection methods, and evaluated the mean and standard deviation of AUC and AUPR of each model on 5 different 10-fold cross validation partitions of the data. As shown in Table 1, all the feature selection algorithms can significantly improve the AUC and AUPR of logistic regression, where the multivariate filter method (CFS) and the wrapper method achieved better prediction performance than the univariate filter and embedded methods on our dataset. Besides, when using multivariate filter, wrapper and embedded selection methods, combining knowledge-based features in predictive modeling can also improve the prediction performance to some extent.

To evaluate the performance of different supervised learning algorithms, we built different learning models on various feature sets (the combination of original and knowledge-based features were used as candidate features in this and the following experiments). As shown in Table 2, GLM methods (logistic regression and Cox) achieved the best performance on our dataset after performing feature selection. Naïve Bayes also got decent AUC and AUPR and their trends are similar with GLM. The decision tree method CART did not work well on our dataset. In comparison, random forest achieved the best performance when applied on all candidate features, but its performance cannot be stably increased by feature selection (except wrapper).

We also compared the performance of our approaches to the state-of-the-arts risk models: Framingham and CHA₂DS₂-VASc scores. We randomly split the dataset to a training set with 60% instances and a testing set with remained instances, and selected feature engineering and supervised learning algorithms based on the above experiments to train our prediction models on the training set. Then we applied both previous models and our trained models on the same testing set, and computed the AUC and AUPR of each model. This process was repeated 5 times, and the means and standard deviations of AUC and AUPR of the models were compared. As shown in Figure 4, the prediction performance of our models outweigh the Framingham and CHA₂DS₂-VASc models.
Discussion

In this study, we compared several feature selection and supervised learning methods in building TE prediction models for AF patients. The GLM (logistic regression and Cox) and Naïve Bayes methods did not work well on the whole feature set, but their prediction performance can be significantly improved by appropriate feature selection. This is probably because both GLM and Naïve Bayes have the assumption of no multicollinearity between the features, but the whole feature set is highly redundant and intercorrelated, which negatively affects the prediction performance of GLM and Naïve Bayes. The feature selection algorithms, especially the CFS algorithm that minimizes the intercorrelation and the wrapper algorithm that directly optimizes the AUC, can reduce the redundancy of features and therefore increase prediction performance. As a whole, the decision tree methods (CART and random forest) did not achieve satisfactory performance on our dataset, probably due to the problem of over-fitting. However, by implicitly embedding feature selection, random forest worked relatively well on the whole feature set with high intercorrelation. Besides, in this study, the wrapper selection method achieved the best prediction performance in terms of AUC and AUPR, because the method directly optimizes the performance metric of the specific learning models. However, the time complexity of wrapper selection is very high, which is not practicable for larger datasets. In addition, because the knowledge-based combination features provided by clinicians can describe high-level clinical concepts, adding them during feature engineering can increase the prediction performance while reducing model complexity, when appropriate feature selection is performed.
Table 3. Risk factors in different models

<table>
<thead>
<tr>
<th>Framingham score</th>
<th>CHA2DS2-VASc score</th>
<th>Commonly selected risk factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Prior stroke/TIA</td>
<td>Prior TE</td>
<td>Prior TE</td>
</tr>
<tr>
<td>Sex</td>
<td>Congestive heart failure</td>
<td>Ischemic stroke confirmed by CT or MRI</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>Hypertension</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>Diabetes mellitus</td>
<td>Left ventricular posterior wall thickness</td>
</tr>
<tr>
<td></td>
<td>Vascular disease</td>
<td>Left ventricular ejection fraction</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>Total cholesterol</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Myocardial infarction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intracranial hemorrhage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Drug use for ventricular rate control</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Years since last TIA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Years since diabetes diagnosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Years since paroxysmal supraventricular tachycardia</td>
</tr>
</tbody>
</table>

In addition to achieving higher predictive performance than existing TE prediction models for AF patients, our approach also identified potential risk factors that had not been commonly used. Table 3 shows the risk factors in the previous Framingham and CHA2DS2-VASc models, as well as the factors that were most commonly selected by multiple feature selection methods in this study. The identified risk factors and their odds ratios in logistic regression were verified by clinicians, concluding that the majority of new risk factors, including cardiovascular problem histories, disease durations, relevant electrocardiography and laboratory tests, as well as medications, are interpretable and reasonable to clinicians.

Despite the promising results in building TE prediction models for AF patients, there are still several aspects of the approach that could be improved. First of all, in this study, we combined knowledge-based features in feature construction, improving the performance of the resulting models. Besides this, there are also other forms of domain knowledge could be used to enhance the predictive power and/or interpretability of models. For example, clinicians have some knowledge and common sense about the impact of features on a target outcome, from their experience or literature. These knowledge could be built as constraints in the modeling algorithms, which could reduce the bias of a specific dataset and then improve the applicability of the models.

Secondly, we applied the state-of-the-arts feature selection algorithms in machine learning to identify risk factors and used performance metrics AUC and AUPR to evaluate their performance. However, the statistical significance (p-value) of the factors, which is critical for a model to be published and adopted in real clinical practices, was not considered. Though the traditional stepwise selection methods can ensure the significance of selected factors, their prediction performance is usually not good enough. Therefore, a new wrapper-based algorithm that optimizes both prediction power and statistical significance would be a practical method to build more interpretable models.

Thirdly, in this study, we only used original features of CAFR data and knowledge-based combinations as candidate features for building prediction models. There are some frequent pattern mining and pattern abstraction methods could be used to discover more complex co-occurrence or temporal patterns, which could be used as combination features in building more effective and interpretable prediction models.

Lastly, in order to use our prediction models in practice, we are also developing a mobile phone app which is named Health Risk Advisor. The TE prediction models can be integrated into the app to provide risk assessment to AF patients, and alert physicians once their patients’ risk levels are changed. A clinical pilot trial for the prediction models could also be conducted in the future based on Health Risk Advisor.
Conclusion

AF significantly increases the risk of ischemic stroke and TE, and accurate prediction of TE for AF patients is critical for early intervention and prevention. In this study, we used integrated machine learning approaches, including data curation, feature engineering and supervised learning, to build TE prediction models for AF patients from CAFR data. The experimental results show that our approach can achieve significantly better prediction performance than previous TE risk models for AF, and identify new potential risk factors as well.

References

Data-Driven Prediction of Beneficial Drug Combinations in Spontaneous Reporting Systems

Ying Li, PhD, Ping Zhang, PhD, Zhaonan Sun, PhD, Jianying Hu, PhD

Center for Computational Health, IBM T.J. Watson Research Center, NY, USA

Abstracts

Post-market withdrawal of medications because of adverse drug reactions (ADRs) could result in loss of effective compounds which are effective for treating a specific disease but have unfavorable benefit-to-harm ratio. Recent therapeutic successes have renewed interest in drug combinations, which could work synergistically to improve therapeutic efficacy or work antagonistically to alleviate the risk of the ADRs. However, experimental screening approaches are costly and often can identify only a small number of drug combinations. Spontaneous reporting systems (SRSs) routinely collect adverse drug events (ADEs) from patients on complex combinations of medications and provide an empirical and cost-effective source to detect unexpected beneficial drug combinations. In this paper, we proposed a novel data-driven method for the prediction of drug combinations where one drug could reduce the ADRs of the other, based on data from SRSs. The predictive model was shown to be effective using a commonly used evaluation approach in pharmacovigilance by constructing a known drug-drug interaction (DDI) reference standard. The method was applied to perform large-scale screening on SRS data for drug-ADR-drug triples where polypharmacy could potentially reduce the ADR. Analysis of the top ranking candidates showed high level of clinical validity.

Introduction

Combinatorial drug therapy is useful for combating complex and refractory diseases such as acquired immune deficiency syndrome (AIDS)\(^1\), cancer\(^2\) and Type 2 diabetes mellitus (T2DM)\(^3\). Two main reasons for using combinatorial drug therapy are that the combination of drugs could work synergistically to improve therapeutic efficacy or work antagonistically to alleviate the risk of adverse drug reactions. A prominent example is the combined use of aspirin and dipyridamole has been shown to be more beneficial and safer than using either of the drugs alone for secondary prevention of stroke in the United States\(^4\).

Despite the increasing number of drug combinations in use, there is significant challenge in discovering beneficial drug-drug combinations (DDCs) in a scalable manner. To date most of them were found in the clinic through experience, or experimentally derived by dose-response curves for each pair of drugs against a protein target\(^5,6\). In recent years, various sources of large-scale data on drugs are emerging, which involve detailed chemical, pharmacological, and pharmaceutical data along with sequence, structure, and pathway information about drug targets\(^7-9\). The systematic identification of novel DDCs is expected to contribute to the development of combinatorial drug therapy\(^6,10,11\). However, the aforementioned studies are based on in-vitro experimental data, data collected from limited number of participants in clinical trials, or from limited well-known drug combinations extracted from FDA orange book\(^12\). Spontaneous reporting systems (SRSs) routinely collect drug-induced adverse drug events (ADEs) from patients on single medication or complex combinations of medications, which provide an opportunity to discover unexpected beneficial drug combinations for ADE reduction.

Many researchers have used SRSs to identify drug combinations that lead to unanticipated harmful adverse events, hereinafter referred to as drug-drug interactions (DDIs), and developed methodologies to effectively mine this database. For instance, Noren et al.\(^13\) implemented a three-way disproportionality measure to identify suspected DDIs, and evaluated the method using empirical examples among the 20 highest predictions. Harpaz et al.\(^14\) used association rule mining to identify multi-item ADE associations from the FDA adverse event reporting systems (FAERS), and 4% of results were characterized and validated as DDIs by an expert. Tatonetti et al.\(^15\) mined FAERS reports for side-effect profiles related to glucose homeostasis and uncovered a novel interaction between pravastatin and paroxetine that causes a potentially hazardous increase in blood glucose levels. Theoretically, we should be able to identify not only adverse drug combinations (DDIs) but also beneficial drug combinations (DDCs) for ADE reduction from SRSs. Zhao et al.\(^16\) proposed the first of its kind to identify DDCs based on FAERS. Specifically, they utilized difference-in-differences estimators\(^17\) to look for drug pairs in which Drug B, when taken with Drug A, could reduce reports of adverse events from patients taking Drug A. Consequently they discovered that the
combined therapy of *rosiglitazone* and *exenatide* was able to reduce the reported incidences of *myocardial infarction* associated with the use of *rosiglitazone* alone. One known issue with SRS data is that selection biases resulting from the nonrandom selection of subjects exposed to the drug and experiencing adverse events could lead to many false positive associations between the drug and the ADR when a causative covariate (for instance, a patient’s disease state or other medications) is not accounted for. In order to alleviate such bias, Zhao *et al.* 16 conducted an extensive post-hoc analysis based on stratification of the data on predefined covariates such as the use of *metformin* and Type 2 diabetes. However, such a stratification requires significant domain knowledge and sometimes enumeration of important covariates, which is intractable for any large-scale analysis. The propensity score matching (PSM) is the most developed and popular strategy for causal analysis in observational studies 18 that yield an unbiased estimate of treatment effects. Tatonetti *et al.* 5 successfully applied the PSM for single drug-ADR detection and results showed that PSM can reduce selection bias associated with drugs reported in the case reports and therefore decrease the false positive associations.

In this study, we propose a novel data-driven method, which incorporates PSM, for large-scale prediction of drug combinations where one drug could reduce the ADRs of the other, based on FAERS data. We denote Drug A as a medication that could cause a specific ADR and Drug B as a medication that could reduce the reported rate of Drug A-induced ADR. We hypothesize that Drug B can alter the chance of developing Drug A-induced ADR based on two mechanisms: the additive or subtractive individual effect of Drug B itself, and the interaction effect with Drug A. Our method is able to predict the combined effect, which takes both mechanisms into consideration. In addition, since both beneficial and harmful drug interactions have gained much interest within the research community, we also report results associated with the interaction effect alone. The knowledge (ground truth) of beneficial drug combinations in terms of ADR reduction is scarce if not nonexistent, we therefore follow a common practice in pharmacovigilance and evaluate the proposed method based on a reference standard of known DDIs and their related ADRs in our experiments.

**Materials and Methods**

**Data sources**

FAERS: FDA adverse event reporting systems

FAERS is a spontaneous reporting system maintained by U.S. Food and Drug Administration (FDA) 19. FAERS contains case reports of suspected ADRs, which are either obligatorily submitted by pharmaceutical companies, or are voluntarily reported by healthcare professionals and consumers. Drugs are entered in a report using free text, which can be brand or generic names, while suspected ADRs are coded using MedDRA terms 20. In addition, the reporters sometimes link the medications to their indications, which are also coded using MedDRA terms. In order to gain statistical power, we normalized drug names to their chemical compounds using the STITCH database 21, which maintains synonym lists for chemicals, and relationships between drugs and their chemical compounds. For example, *quinapril hydrochloride* and *hemokvin* are mapped to the main ingredient *quinapril*.

SIDER: a Side Effect Resource

The SIDER (SIDe Effect Resource) 22 is a publicly available resource that relates the medications to their known side effects, or ADRs. The relationships are extracted by an automatic method from the FDA Structured Product Labels. Presently, the medications are coded using STITCH and the ADRs are coded using MedDRA terms. For instance, the medication *lansoprazole* (CID000216416) is described to cause the MedDRA-coded side effect *gastrointestinal pain*. For this study, we utilized the latest version SIDER 4.1.

DrugBank: a comprehensive knowledgebase for drugs, drug actions, and drug targets

DrugBank 23 is a knowledge base that contains detailed biomedical and pharmacological information about drugs, their mechanisms and their targets. In addition, DrugBank provides a set of 12,128 ingredient level drug- drug interactions (DDIs), and most of which are along with a brief textual description of the interaction. For example, an interaction between *nalidixic acid* and *warfarin* is described as “*nalidixic acid* may increase the anticoagulant effect of *warfarin*”. Another example is “the combined use of *amiodarone* and *lovastatin* increase the risk of severe myopathy/rhabdomyolysis”, which links the DDI between *amiodarone* and *lovastatin* to its ADRs of myopathy or rhabdomyolysis that can be mapped to MedDRA.
terms. The version of DrugBank used in this study is 4.0.

Methods

The overall workflow of the study is shown in Figure 1, which consisted of three steps: (1) Identify and filter potential candidate triples of Drug A-ADR-Drug B. (2) Compute propensity scores for individual drugs. (3) Determine the associations for each specific Drug A-ADR-Drug B triple.

![Figure 1. An overview workflow of the study.](image)

Step 1. Filtering of potential candidate triples of Drug A-ADR-Drug B

Three criteria in combination with a statistical method were used to limit the search space of the set of Drug A-ADR-Drug B triples. From all pairs of Drug A-ADR that were reported in the FAERS case report data for more than 200 times (Criteria 1), we identified the candidate pairs of Drug A-induced ADR based on the observed Odds Ratio (OR). Let Drug denote a Drug A, the OR of Drug developing an ADR is defined as follows:

$$
\text{OR}_{\text{drug}} = \frac{\text{odds}(\text{ADR} = 1|\text{drug} = 1)}{\text{odds}(\text{ADR} = 1|\text{drug} = 0)}
$$

(1)

where \( \text{odds}(X) = \frac{\Pr(x)}{1 - \Pr(x)} \).

An OR greater than 1 indicates the chance that Drug develops the ADR is higher than the expected background rate, and an OR smaller than 1 indicates the chance that Drug develops the ADR is lower than the expected background rate. By replacing the probabilities in (1) with empirical probabilities, we obtained the observed OR for each Drug A-ADR pair. We then used Chi-Square test to determine whether the observed OR for a drug-ADR pair was significantly greater than 1. Any pair with p-value less than 0.05 were selected into the initial set of Drug A-ADR pairs. We further filtered out those pairs of Drug A-ADR that were not mentioned in the SIDER 4.1 (Criteria 2).

For each Drug A-ADR pair, we selected medications as Drug Bs that are co-prescribed with Drug A for a minimum of 500 case reports, regardless of the association between a Drug B and the ADR of interest (Criteria 3). We denoted the set of all candidate Drug A-ADR-Drug B triples as \( C = \{(\text{Drug A, ADR, Drug B})\} \), and elements in \( C \) is denoted as \( C_i \). The associations between Drug B and the ADR in each \( C_i \), when
patients are prescribed with Drug A will be our interest throughout this paper.

Step 2. Compute propensity scores for individual drugs

The propensity score is the conditional probability of being exposed to a drug, given a set of baseline characteristics. Using propensity scores, potential bias due to treatment selection could be mitigated and the response being evaluated is conditionally independent given the measured baseline characteristics. To adjust potential selection bias, we calculated propensity score for each individual drug that is either Drug A or Drug B in C.

For each individual drug, we selected the baseline characteristics as the top 200 most relevant medications and indications. The relevance of a medication or an indication to a drug is measured by the phi correlation coefficient. We use Rx and Dx to denote the relevant medications and indications respectively.

To obtain the propensity score, a logistic regression is performed for each drug, which is expressed as the following.

\[
\logit(P(Drug = 1)) = \alpha + \sum_{i=1}^{200} \delta_i Rx_i + \sum_{j=1}^{200} \gamma_j Dx_j
\]

(2)

Where \(\logit(x) = \log(x/(1-x))\), \(\alpha\), \(\delta\) and \(\gamma\) are the coefficients in the logistic regression. Once we obtain the estimates of the coefficients, the propensity score is estimated as the predicted probability of receiving the drug for each case report, i.e. predicted value of \(P(Drug = 1)\).

Step 3. Determine the association for the Drug A-ADR-Drug B triple

Within each triple \(C_i = (\text{Drug A}, \text{ADR}, \text{Drug B})\), it has been established in Step 1 that the chance that Drug A develops the ADR of interest is higher than the expected background rate. By adding the second medication - Drug B, the chance of developing the ADR of interest would possibly alter.

Specifically, we are interested in two predicted scores. First, given that the patient is prescribed with Drug A, whether adding Drug B will modify the chance of developing the ADR of interest. Second, given that the patient is prescribed with both Drug A and Drug B, whether Drug B will interact with Drug A to alter the chance of developing the ADR of interest. These two predicted scores can be generated by a regularized logistic regression as follows:

\[
\logit(P(ADR = 1)) = \beta_0 + \beta_1 DrugA + \beta_2 P_1 + \beta_3 DrugB + \beta_4 P_2 + \beta_5 DrugA * DrugB + \lambda |\beta|_1
\]

(3)

Where \(P_1\) and \(P_2\) are the propensity scores for Drug A and Drug B, respectively. \(\beta_i\) for \(i = 0, \ldots, 5\) are the coefficients, \(\beta = (\beta_1, \ldots, \beta_5)\) denote the vector of linear coefficients, and \(|\beta|_1\) denote the L-1 norm. The last term on the right hand side of Formula (3) is the least absolute shrinkage and selection operator (LASSO) regularizer, which could enforce estimates of \(\beta\) to be sparse, and \(\lambda\) is the tuning parameter of the regularizer. In this paper, for each \(C_i = (\text{Drug A}, \text{ADR}, \text{Drug B})\), we selected an optimal \(\lambda\) by 3-fold cross-validation. \(P_1\) and \(P_2\) are obtained from Step 2. The purpose of adding propensity scores in the logistic regression is to mitigate the drug selection bias given other patient characteristics such as medications that patients on and underlying diseases that patients have.

The value of \((\beta_1 + \beta_2)\) is referred to as predicted beneficial score for taking Drug B. The predicted beneficial score specifies the degree that a patient who is on Drug A could benefit or suffer from taking Drug B for the ADR of interest. A predicted beneficial score smaller than 0 indicates that patients who are on Drug A could benefit from taking Drug B in terms of reducing the ADR of interest, and a predicted beneficial score greater than 0 indicates that patients who are on drug A cannot benefit from taking drug B in terms of reducing the ADR of interest. The value of \(\beta_5\) is referred to as predicted interaction score. A predicted interaction score smaller than 0 indicates an antagonistic interaction between Drug A and Drug B which decrease the chance of developing the ADR of interest, while a predicted interaction score greater than 1 signifies a synergistic interaction between Drug A and Drug B which increase the chance of developing the ADR of interest. The later scenario is also referred to as DDIs.

Evaluation

The aim of this study is to identify a set of Drug Bs that could reduce the reporting rate of Drug A-induced ADRs. The appropriate evaluation of the proposed system should be based on a reference standard
consisted of a list of Drug Bs that are known to or not known to reduce the Drug A-induced ADRs, however, this kind of knowledge is scarce if not nonexistent. In order to demonstrate the effectiveness of the proposed method quantitatively, we assessed the predicted interaction score ($\beta_S$) against a set of known DDIs and their ADRs. Following common practice in pharmacovigilance\textsuperscript{25,26}, we constructed a reference standard, which consisted of DDIs known to cause or not known to cause either of two serious ADRs: rhabdomyolysis and QT prolongation. The DDIs, known to be associated with either rhabdomyolysis or QT prolongation, were extracted from DrugBank knowledgebase that comprise positive controls in the reference standard. For negative controls, we randomly paired drugs in the FAERS, excluded those that are known to interact with each other for any ADR, and then randomly selected the same number of negative controls from the rest of drug pairs. The rationale of selecting equal representation of positive and negative controls is to avoid class imbalance that can bias performance metrics to favor a method that is calibrated towards true positive rates when a reference set predominantly contains positive controls or specificity (1-\text{false positive rate}) when negative controls predominate\textsuperscript{27,28}. To quantitatively assess the performances of the proposed method, we generated receiver operating characteristic (ROC) curves, which are graphical plots of true positive rate vs. false positive rate. The whole ROC curve can be plotted by varying the threshold value or prediction score, above which the output is predicted as positive and negative otherwise. We also calculated the area under ROC curve (AUC) in our experiments.

Results

Data characteristics

FAERS data were collected from January 2004 to September 2015 accounting for 6,434,615 case reports, which contain 9,315 unique STITCH coded drugs and 18,159 MedDRA coded unique ADEs. On average, each report contains 3.5 different generic drugs, 2.2 different indications and 3.4 different ADEs.

System Evaluation

Table 1 shows statistics of and examples of positive and negative controls in the reference standard for DDIs and their ADRs. The reference standard was based on DrugBank knowledgebase. For instance, according to DrugBank knowledgebase, the concurrent use of cisapride and tacrolimus may result in increased risk of QT-interval prolongation. In total, 50 positive controls and negative controls were generated for rhabdomyolysis, and 61 positive controls and negative controls were created for QT prolongation. We further defined the relevant MedDRA preferred term for two ADRs, which were rhabdomyolysis and electrocardiogram QT prolonged respectively.

Figure 2 shows the resulting ROC curves for the classification of 100 drug pairs for rhabdomyolysis and 122 drug pairs for QT prolongation as true/false DDI using the proposed method. The AUC is 0.80 and 0.70 for rhabdomyolysis and QT prolongation respectively (1 indicates perfect classifier and 0.5 indicates random classifier), which indicates that the proposed system is effective for identifying DDIs and their ADRs. We then utilized the method to identify drugs (Drug Bs) that could reduce the ADRs of the other drugs (Drug As).

Table 1. Statistics and examples of positive and negative controls of the reference standard for DDIs.

<table>
<thead>
<tr>
<th>ADR</th>
<th>Rhabdomyolysis</th>
<th>QT prolongation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedDRA preferred term used in FAERS</td>
<td>RHABDOMYOLYSIS</td>
<td>ELECTROCARDIOGRAM QT PROLONGED</td>
</tr>
<tr>
<td>Number of positive controls</td>
<td>50</td>
<td>61</td>
</tr>
<tr>
<td>Examples of the combined use of drugs associated with the ADR</td>
<td>atorvastatin_repaglinide</td>
<td>cisapride_tacrolimus</td>
</tr>
<tr>
<td></td>
<td>simvastatin_amiodipine</td>
<td>methadone_tripipramine</td>
</tr>
<tr>
<td></td>
<td>colchicine_fluvastatin</td>
<td>fluconazole_voriconazole</td>
</tr>
<tr>
<td>Number of negative controls</td>
<td>50</td>
<td>61</td>
</tr>
<tr>
<td>Examples of the combined use of drugs not known to interact</td>
<td>losartan_torsemide</td>
<td>sotalol_valsalan</td>
</tr>
<tr>
<td></td>
<td>pravastatin_electrolytes</td>
<td>levofloxacin_doxorubicin</td>
</tr>
<tr>
<td></td>
<td>atorvastatin_hydmorphone</td>
<td>hydroxyzine_haloeperdol</td>
</tr>
</tbody>
</table>
Large-scale prediction of novel drug combinations and their ADR reduction

After establishing the effectiveness of the approach as described above, we proceeded to perform comprehensive predictions of Drug Bs that can decrease the reporting rates of the ADRs induced by Drug As. A candidate set of 694,957 Drug A-ADR-Drug B triples was selected in Step 1 of the method, which involves 449 Drug As, 895 ADRs and 947 Drug Bs. Among them, 322,417 Drug A-ADR-Drug B triples have predicted beneficial scores smaller than 0 indicating that polypharmacy could potentially reduce the ADR.

As mentioned in the method section, the predicted beneficial score ($\beta_3$+$\beta_5$) indicates the degree that a patient who is on Drug A could benefit or suffer from taking Drug B for the ADR of interest, and the predicted interaction score ($\beta_5$) specifies the degree that the interaction effect between Drug B and Drug A on the ADR. By using the cutoff of 0, Table 2 shows the statistics of drug effects on the ADRs of interest computed from the regularized regression model. After adjusting for patient characteristics, 95.6% of Drug A-ADR pairs were positively associated indicating that Drug A increases the reported rate of ADRs, 3.3% of Drug A-ADR pairs were not associated indicating that Drug A does not either increase or decrease the reported rate of ADRs and 1.1% of Drug A-ADR pairs were negatively associated indicating that Drug A decreases the reported rate of ADRs. Moreover, among all Drug A-ADR-Drug B triples, 46.4% of predicted beneficial scores were smaller than 0 indicating that taking Drug B could decrease the chance of developing the ADR induced by Drug A; and 38.5% of predicted interaction scores were smaller than 0 signifying that Drug B could interact antagonistically with Drug A to reduce the ADR.

Table 2. Statistics of the drug effects on the ADR

<table>
<thead>
<tr>
<th>The effect of drug on the ADR (coefficient)</th>
<th>Increase the incidence of the ADR (coefficient &gt; 0)</th>
<th>No association with the ADR (coefficient = 0)</th>
<th>Decrease the incidence of the ADR (coefficient &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A ($\beta_1$)</td>
<td>95.6%</td>
<td>3.3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Predicted beneficial score ($\beta_3$+$\beta_5$)</td>
<td>34.3%</td>
<td>19.3%</td>
<td>46.4%</td>
</tr>
<tr>
<td>Predicted interaction score ($\beta_5$)</td>
<td>14.6%</td>
<td>46.9%</td>
<td>38.5%</td>
</tr>
</tbody>
</table>

We are more interested in negative values of predicted beneficial score, and negative values of predicted interaction score (both of them are highlighted in bold in Table 2). Figure 3 (a) shows the histogram of
numbers of beneficial Drug Bs per Drug A-ADR pair. In total, the reporting rates of 4,950 drug A-induced ADRs can be reduced by at least one Drug B. On average, the reporting rate of a specific Drug A-induced ADR can be decreased by 65 Drug Bs. Figure 3 (b) shows the histogram of numbers of Drug Bs that could interact with Drug A antagonistically to reduce Drug A-induced ADR. In total, the reporting rates of 4,825 drug A-induced ADRs could be reduced by at least one Drug B’s antagonistic interaction with Drug A. On average, the reporting rate of a specific drug A-induced ADR could be decreased by 55 Drug Bs.

Figure 3. Statistics of Drug Bs that reduce the reporting rates of Drug A-induced ADRs. (a) Histogram of Drug Bs with favorable predicted beneficial score per Drug A-ADR pair. (b) Histogram of drug Bs that interacts antagonistically with Drug A per Drug A-ADR pair.

Table 3. List of 15 predicted beneficial drug combinations and their ADR reduction

<table>
<thead>
<tr>
<th>Drug A name</th>
<th>ADRs associated with drug A</th>
<th>Drug B name</th>
<th>Predicted beneficial score</th>
<th>Common ATC code</th>
<th>Evidence for combined use</th>
</tr>
</thead>
<tbody>
<tr>
<td>benazepril</td>
<td>DIZZINESS</td>
<td>amlodipine besylate</td>
<td>-0.57</td>
<td>yes</td>
<td>F</td>
</tr>
<tr>
<td>atovaquone</td>
<td>PYREXIA</td>
<td>proguanil</td>
<td>-0.36</td>
<td>yes</td>
<td>F</td>
</tr>
<tr>
<td>rofecoxib</td>
<td>MYOCARDIAL INFARCTION</td>
<td>pamidronate</td>
<td>-0.33</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>rosiglitazone</td>
<td>MYOCARDIAL INFARCTION</td>
<td>exenatide</td>
<td>-0.32</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>progesterone</td>
<td>BREAST CANCER</td>
<td>adalimumab</td>
<td>-0.27</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>trimethoprim</td>
<td>PYREXIA</td>
<td>sulfamethoxazole</td>
<td>-0.17</td>
<td>yes</td>
<td>F</td>
</tr>
<tr>
<td>exemestane</td>
<td>ARTHRALGIA</td>
<td>everolimus</td>
<td>-0.16</td>
<td>yes</td>
<td>III</td>
</tr>
<tr>
<td>amoxicillin</td>
<td>DIARRHOEA</td>
<td>clavulanic acid</td>
<td>-0.15</td>
<td>yes</td>
<td>IV</td>
</tr>
<tr>
<td>ampicillin</td>
<td>PYREXIA</td>
<td>sulbactam</td>
<td>-0.15</td>
<td>yes</td>
<td>F</td>
</tr>
<tr>
<td>desmopressin</td>
<td>HYPONATRAEMIA</td>
<td>somatropin</td>
<td>-0.15</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>sertraline</td>
<td>ANXIETY</td>
<td>nicotinic acids</td>
<td>-0.14</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>sumatriptan</td>
<td>MIGRAINE</td>
<td>naproxen</td>
<td>-0.14</td>
<td>no</td>
<td>F</td>
</tr>
<tr>
<td>olanzapine</td>
<td>DIABETES</td>
<td>biperiden</td>
<td>-0.13</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>clindamycin</td>
<td>DIARRHOEA</td>
<td>benzoil</td>
<td>-0.13</td>
<td>yes</td>
<td>F</td>
</tr>
<tr>
<td>fluticasone</td>
<td>DYSPNOEA</td>
<td>salmeterol</td>
<td>-0.13</td>
<td>yes</td>
<td>F</td>
</tr>
</tbody>
</table>

F: FDA approved drug combination; III: phase III clinical trial; IV: phase IV clinical trial
We further analyzed top 15 predictions for two scores. Table 3 displays a list of the top 15 Drug A-ADR-Drug B triples sorted with our predicted beneficial scores wherein the Drug B most dramatically reduced Drug A-induced ADR. In parallel, Table 4 presents a list of the top 15 Drug A-ADR-Drug B triples sorted with our predicted interaction scores wherein the drug B results in the best reduction of drug A-induced ADR. We investigated whether the predicted Drug B were prescribed with Drug A for the same disease using ATC code. Twelve Drug A-Drug B pairs in Table 3 and nine Drug A-Drug B pairs in Table 4 shared at least one ATC code signifying that the predicted drug combinations are useful in clinical practice since drug combinations are usually prescribed to treat or manage the same disease. Finally, the FDA orange book and clinical trials were interrogated to check clinical validity of the top ranked drug combinations. Nine drug combinations in Table 3 and two drug combinations in Table 4 have been validated in the orange book or investigated in the clinical trials, indicating their clinical validity.

**Table 4. List of 15 predicted interacted drug combinations and their ADR reduction**

<table>
<thead>
<tr>
<th>Drug A name</th>
<th>ADR associated with drug A</th>
<th>Drug B name</th>
<th>Predicted interaction score</th>
<th>Common ATC code</th>
<th>Evidence for combined use</th>
</tr>
</thead>
<tbody>
<tr>
<td>niacin</td>
<td>FLUSHING</td>
<td>vitamin e</td>
<td>-0.34</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>rofecoxib</td>
<td>MYOCARDIAL INFARCTION</td>
<td>pamidronate</td>
<td>-0.31</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>rosiglitazone</td>
<td>MYOCARDIAL INFARCTION</td>
<td>exenatide</td>
<td>-0.31</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>progesterone</td>
<td>BREAST CANCER</td>
<td>adalimumab</td>
<td>-0.27</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>telaprevir</td>
<td>FATIGUE</td>
<td>peginterferon alfa-2b</td>
<td>-0.25</td>
<td>no</td>
<td>III</td>
</tr>
<tr>
<td>lapatinib</td>
<td>DIARRHOEA</td>
<td>vinorelbine</td>
<td>-0.17</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>exemestane</td>
<td>ARTHRALGIA</td>
<td>everolimus</td>
<td>-0.15</td>
<td>yes</td>
<td>III</td>
</tr>
<tr>
<td>desmopressin</td>
<td>HYPONATRAEMIA</td>
<td>somatropin</td>
<td>-0.15</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>clofarabine</td>
<td>FEBRILE</td>
<td>cytarabine</td>
<td>-0.13</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>finasteride</td>
<td>SEXUAL DYSFUNCTION</td>
<td>zoledronic acid</td>
<td>-0.12</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>liraglutide</td>
<td>NAUSEA</td>
<td>pioglitazone</td>
<td>-0.12</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>midazolam</td>
<td>CARDIAC ARREST</td>
<td>pancuronium</td>
<td>-0.11</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>olanzapine</td>
<td>DIABETES MELLITUS</td>
<td>levomepromazine</td>
<td>-0.11</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>cisplatin</td>
<td>RENAL FAILURE ACUTE</td>
<td>mannitol</td>
<td>-0.10</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>melphalan</td>
<td>NEUTROPENIA</td>
<td>granisetron</td>
<td>-0.09</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

III: Phase III clinical trial

**Discussions**

Results in Figure 2 demonstrated that the proposed method is effective in identifying and prioritizing DDIs with their ADRs. By comparing Table 3 with Table 4, more clinical evidences were found for drug pairs when the overall beneficial effect of Drug B was considered. Typically, drug combinations are developed to aim at a more effective treatment than the monotherapy for a single disease such as the combined use of benazepril and amlodipine besylate, ranked 1 in Table 3, for treating hypertension. And the combined use of atovaquone and proguanil, ranked 2 in Table 3, for treating or preventing malaria. However, very few clinical trials focus on evaluating the reduction of the Drug A-induced ADRs that the Drug B is able to achieve. Our method, in combination with a spontaneous reporting system, showed that the use of amlodipine besylate reduced benazepril-induced ADRs such as angioedema, cough and dizziness, with the most reduction in dizziness. Similarly, results showed that the use of proguanil reduced atovaquone-induced ADRs such as diarrhea, vomiting and pyrexia, with the most reduction in pyrexia. Cisplatin-acute renal failure-mannitol triple in Table 4 is an example that mannitol is used to reduce the chance of acute
renal failure (ARF) with the use of cisplatin. Although a couple of studies have reached to different conclusions about additional use of mannitol, either leads to reduction in the incidence of nephrotoxicity or is associated with the null result \(^{29,30}\). Our results showed that mannitol could reduce ARF through an antagonistic interaction with cisplatin.

Similar to DDI detection task wherein very few of publicly available data sources linking the potential DDIs to their ADRs \(^{31}\), very few of mined DDC data sources that link DDCs to their beneficial clinical effects. For example, the study conducted by Iwata et al. \(^{11}\) only predicted that the combined use of adrenaline and chloroprocaine was beneficial without relating them to a specific clinical effect. Our predicted data set is capable to relate DDCs to their clinical effects.

**Limitations and Future Research**

In Table 2, the use of Drug B may sometimes increase the reporting rate of Drug A-induced ADRs (34.3% of all triples). Specifically, there is a scenario where a Drug B that could decrease some Drug A-induced ADRs while increase the other Drug A-induced ADRs. For example, although additional use of sulfamethoxazole decreased trimethoprim-induced pyrexia, ranked 6 in Table 4, it also increased trimethoprim-induced hypersensitivity. In future work, a more comprehensive assessment should be conducted to evaluate the comprehensive effect of Drug B among patients who are on Drug A.

More legitimate cutoff values for the predicted beneficial and interaction scores, rather than 0 used in the study, could make our proposed approach more practical for supporting domain experts. Due to the heterogeneity in the mechanism of drug combination, as well as the heterogeneity in the sample sizes, a universal cutoff value may not be available. In our future research, we will explore statistical methods to determine specific cutoff values for different drug combinations.

The full database screening of FAERS to identify all Drug Bs to reduce the reporting rate of Drug A-induced ADRs are very time consuming simply because they are severely dependent on the searches of a vast space of possible drug pairs and relevant ADRs. In order to reduce the search space, we utilized the following criteria in the study: (1) Drug A should have more than 200 reports with the ADR; (2) Drug A-ADR signals should have been corroborated in SIDER 4.1 and (3) Drug A and Drug B are mentioned for a minimum of 500 case reports. The values in criteria (1) and (3) were also utilized in Zhao et al.’s study \(^{16}\). In future work, we plan to implement the algorithm in the high performance-computing environment so that we can relax these values in criteria to obtain more beneficial drug combinations and their clinical effects.

**Conclusion**

In this study, we proposed a novel data-driven method for large-scale prediction of drug combinations where one drug could reduce the ADRs of the other, based on FAERS data set. We validated the method against a known DDI reference standard, and applied the method to perform large-scale screening on FAERS data for drug-ADR-drug triples where polypharmacy could potentially reduce the ADR. Analysis of the top ranking candidates demonstrated high level of clinical validity and the usefulness of the approach.

**References**


22. Kuhn M, Campillos M, Letunic I, Jensen LJ, and Bork P. A side effect resource to capture phenotypic effects of drugs. Molecular systems biology 2010;6:


CMedTEX: A Rule-based Temporal Expression Extraction and Normalization System for Chinese Clinical Notes

Zengjian Liu, MS1, Buzhou Tang, PhD1*, Xiaolong Wang, PhD1, Qingcai Chen, PhD1, Haodi Li, MS1, Junzhao Bu, MS1, Jingzhi Jiang, MS1, Qiwen Deng, MS2, Suisong Zhu, MS2
1Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China; 2The Sixth People’s Hospital of Shenzhen, Shenzhen, China

Abstract

Time is an important aspect of information and is very useful for information utilization. The goal of this study was to analyze the challenges of temporal expression (TE) extraction and normalization in Chinese clinical notes by assessing the performance of a rule-based system developed by us on a manually annotated corpus (including 1,778 clinical notes of 281 hospitalized patients). In order to develop system conveniently, we divided TEs into three categories: direct, indirect and uncertain TEs, and designed different rules for each category of them. Evaluation on the independent test set shows that our system achieves an F-score of 93.40% on TE extraction, and an accuracy of 92.58% on TE normalization under “exact-match” criterion. Compared with HeidelTime for Chinese newswire text, our system is much better, indicating that it is necessary to develop a specific TE extraction and normalization system for Chinese clinical notes because of domain difference.

Introduction

Time is one of the most important aspects of information as information is usually event-centred and organized in order of time. To understand progress of events, people have to capture temporal information related to them, including temporal expressions (TEs) and temporal relations. Because of this, temporal information extraction as a fundamental task of natural language processing (NLP) has been always attracting a great deal of attention. A large number of systems using various methods, including rule-based and machine learning (ML)-based, have been developed to extract temporal information in text in different languages (e.g., English, Chinese, French, etc.) from different domains such as newswire and social media. In the clinical domain, there also have been a few temporal information extraction systems. However, most of them are designed for English clinical notes from electronic medical record (EMR) systems, and very limited studies have been carried out on Chinese clinical notes.

In recent years, with the rapid growth of EMRs in China, information extraction from Chinese clinical notes has attracted more and more attention of Chinese researchers. In this study, we developed a temporal expression (TE) extraction and normalization system for Chinese clinical notes, assessed its performance on a manually annotated corpus of Chinese clinical notes, and published our system online. To the best of knowledge, this is one of the earliest studies on TE extraction and normalization for Chinese clinical notes, our system is the first publicly available system, and we believe it will provide valuable insights into NLP research in Chinese clinical text.

Background

TE extraction and normalization are integral parts of information extraction and have been extensively studied in multilingual text in multiple domains. The development of related technology was mainly driven by several public challenges such as the TE recognition and normalization (TERN) challenge1 in 2004, the automatic content extraction (ACE) challenges2 in both 2005 and 2007, and a series of TempEval challenges: TempEval-13 in 2007, TempEval-24,5 in 2010 and TempEval-36 in 2013. In the early challenges (e.g., the 2004 TERN challenge and the 2007 ACE challenge), only English text in newswire domain was considered. Subsequently text in other languages such as Chinese was also gradually included, and the text was not limited to newswire text. For example, extracting and normalizing TEs in Chinese text from broadcast news, newswire and weblogs was added as a new subtask of the 2005 ACE challenge. In these challenges, both rule-based and ML-based approaches have been investigated, and the rule-based ones almost always showed better performance than the ML-based ones. For example, in the 2010 TempEval-2 challenge, the best TE extraction and normalization system (i.e., HeidelTime7-9) was a rule-based system which significantly outperformed other state-of-the-art ML-based systems such as two conditional random

* Corresponding author; Email: tangbuzhou@gmail.com
fields (CRF)-based systems: TIPSem\textsuperscript{10} and TRIOS\textsuperscript{11}. In the rule-based system, not only word, but also syntactic and semantic information such as part-of-the-speech and semantic role of words were used for rule design\textsuperscript{10, 12, 13}.

In the clinical domain, with the development of clinical NLP, a number of attempts have been proposed for English clinical notes. The early representative clinical NLP systems such as ConText\textsuperscript{14}, CNTRO\textsuperscript{15, 16}, MedLEE\textsuperscript{17}, CTakes\textsuperscript{18}, etc. usually include an individual rule-based module for TE extraction and normalization. As there was no benchmark corpus available for TE extraction and normalization system evaluation, it was difficult to compare them. In 2012, the i2b2 (Center of Informatics for Integrating Biology and Bedside) organized a clinical NLP challenge\textsuperscript{19} of temporal information extraction in clinical text, including a subtask of TE extraction and normalization. This challenge provided a benchmark corpus to evaluate different systems developed by researchers all over the world. In this challenge, both rule-based and ML-based methods were proposed for TE extraction and normalization, and finally the rule-based system developed by Sunghwan Sohn et al\textsuperscript{20} achieved the best performance, slightly better than the CRF-based system developed by Yan Xu et al\textsuperscript{21}. Subsequently, the SemEval challenge in 2015 included a clinical TempEval task\textsuperscript{22, 23} including a subtask of TE extraction and normalization in clinical text. Most of systems were rule-based. The information used to design rules included word, pos, clinical entity, etc.

In recent years, with the rapid growth of EMRs in China, the need of information extraction from Chinese clinical notes has become more and more strident, including temporal information extraction. However, very limited studies have been proposed for TE extraction and normalization in Chinese clinical text. Zhou et al\textsuperscript{24} proposed a simple rule-based method to extract and normalize TEs in Chinese clinical text and evaluated the method on a very small manually annotated dataset with only 1,207 TEs from 147 medical records. Despite the important contributions of previous studies on TE extraction and normalization in Chinese clinical text, none has systematically analyzed the difference of the TE extraction and normalization systems for clinical text in Chinese and English, and the difference of the TE extraction and normalization systems for Chinese clinical text in the newswire domain and clinical domain, which is very important for system porting from English clinical text or Chinese newswire text to Chinese clinical text.

In this paper, we analyzed the characteristics of Chinese clinical notes, developed a rule-based TE extraction and normalization system for Chinese clinical text (available at http://icrc.hitsz.edu.cn:8096/CMedTEX), and analyzed which cases were the new cases in Chinese clinical text compared to English clinical text and Chinese newswire text. To the best of our knowledge, this is one of the earliest comprehensive studies on TE extraction and normalization for Chinese clinical notes, and our system is the first publicly available TE extraction and normalization system for Chinese clinical notes, whose source code was also released for application at: http://icrc.hitsz.edu.cn/Article/show/147.html.

**Methods**

**Dataset and annotation**

A real world dataset of 1,778 de-identified clinical notes of 281 hospitalized patients were collected from the sixth people’s hospital of Shenzhen, Shenzhen, China. Two master students with computer science and medical informatics background were recruited to annotate TEs in the clinical notes by following annotation guidelines designed for this study. When creating the annotation guidelines, we referred to the TE annotation guidelines of TimeML\textsuperscript{25, 26} for English newswire text and the 2012 i2b2 NLP challenge\textsuperscript{27} for English clinical text, and defined the same three attributes for each TE, i.e., type, value and modifier, where there were four types (DATE, TIME, DURATION and FREQUENCY) and seven modifiers (APPROX, START, END, MIDDLE, MORE, LESS and NA (default value)), and the value attribute is the standard temporal value defined in ISO8601. Some examples were shown in Figure 1.

To calculate the inter-rater annotation agreement of the two annotators, we asked them to annotate 300 out of 1,778 notes at the same time. When there was any disagreement between the two annotators, a third judge was brought in to select or confirm the different annotations. Following the 2012 NLP challenge\textsuperscript{29, 30}, the inter-rater annotation agreement of the two annotators was calculated under two criteria: “exact-match” and “partial-match”, where “exact-match” and “partial-match” denote that two time expressions are correctly matched if and only if their text spans are exactly the same and overlap with each other respectively. Based on these 300 notes, inter-rater agreements of TE span, type and modifier using Kappa statistics\textsuperscript{29} were 0.9035, 0.9681 and 0.9195, the inter-rater annotation agreements of type value and modifier using accuracy were 0.9756, 0.8773 and 0.9879 under the exact-match criterion (see Table 1), indicating that our annotation was reliable. The remaining 1,478 notes were annotated by one annotator only.
Figure 1. Examples of TEs with attributes. ‘x’ denotes the missing information needing to be determined during normalization.

Table 1. Inter-annotator agreement under both “exact-match” and “partial-match” criteria. “/” denotes there is no available value.

<table>
<thead>
<tr>
<th></th>
<th>Exact-match</th>
<th>Partial-match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>Span</td>
<td>/</td>
<td>0.9035</td>
</tr>
<tr>
<td>Type</td>
<td>0.9756</td>
<td>0.9681</td>
</tr>
<tr>
<td>Value</td>
<td>0.8773</td>
<td>/</td>
</tr>
<tr>
<td>Modifier</td>
<td>0.9879</td>
<td>0.9195</td>
</tr>
</tbody>
</table>

After annotation, we obtained 46,508 TEs. The annotated dataset was randomly divided into training and test sets. The training set consisted of 1,076 records with 28,442 TEs of 170 hospitalized patients, while the test set consisted of 702 records with 18,066 TEs of 111 hospitalized patients. The statistics of the dataset were shown in Table 2 in detail, where “#*” denotes the number of ‘*’. Among four types of TEs, dates accounted for the highest proportion of 36.56% (17,005/46,508), and times accounted for the lowest proportion of 15.84% (7,368/46,508).

Table 2. Statistics of the annotated dataset for TE extraction and normalization.

<table>
<thead>
<tr>
<th></th>
<th>#Patient</th>
<th>#Record</th>
<th>#Temporal expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>170</td>
<td>1,076</td>
<td>10,205 4,500 6,821 6,916 28,442</td>
</tr>
<tr>
<td>Test set</td>
<td>111</td>
<td>702</td>
<td>6,800 2,868 4,214 4,184 18,066</td>
</tr>
<tr>
<td>Total</td>
<td>281</td>
<td>1,778</td>
<td>17,005 7,368 11,035 11,100 46,508</td>
</tr>
</tbody>
</table>

TE extraction and normalization

Although a large number of both machine learning-based and rule-based methods have been proposed for TE extraction, experiments in TempEval-2 and 2012 i2b2 NLP challenges demonstrated that the rule-based methods obviously outperformed the machine learning-base methods, and the best systems of the two challenges were all rule-based4, 19. For TE normalization, almost all systems were rule-based. Therefore, in this study, we investigated the rule-based methods for TE extraction and normalization. In order to develop our system conveniently, we analyzed the characteristics of Chinese clinical notes and classified them into three categories: direct, indirect and uncertain, which denoted complete and exact TEs, incomplete or implicit but exact TEs, and inexact TEs. Figure 2
listed some examples. For each category of TEs, we designed specific rules, which were presented in the following paragraphs in detail.

<table>
<thead>
<tr>
<th>Direct TEs</th>
<th>Indirect TEs</th>
<th>Uncertain TEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011年6月21日 (June 21, 2011)</td>
<td>6月21日 (June 21)</td>
<td>6月中旬 (in the middle of June)</td>
</tr>
<tr>
<td>二零一一年六月二十一日 (June 21, 2011)</td>
<td>朂天 (tomorrow)</td>
<td>10年零 (more than ten years ago)</td>
</tr>
<tr>
<td>2011年12月13日 10:20 (10:20 December 13, 2011)</td>
<td>上个月 (last month)</td>
<td>今晨7:30时 (about 7:30 in the morning)</td>
</tr>
<tr>
<td>二零一一年十二月十三日十点钟 (10 o’clock on December 13, 2011)</td>
<td>3天前 (three days ago)</td>
<td>几个小时前 (several hours ago)</td>
</tr>
<tr>
<td>5个小时 (5 hours)</td>
<td>手术前一天 (the day before surgery)</td>
<td>3至3个半天 (more than 3 hours)</td>
</tr>
<tr>
<td>36周+5天 (36 weeks and 5 days)</td>
<td>12月13号9点 (9 o’clock on December 13)</td>
<td>40-天 (more than 40 days)</td>
</tr>
<tr>
<td>qid (four times per day)</td>
<td>下午四时 (4 p.m.)</td>
<td>2-3天/周 (2 or 3 times per day)</td>
</tr>
<tr>
<td>一日三次 (three times per day)</td>
<td></td>
<td>至少2次/月 (at least 2 times a month)</td>
</tr>
</tbody>
</table>

**Figure 2.** Examples of TEs in the three categories: direct, indirect and uncertain.

1. Direct TEs

It is easy to extract and normalize direct TEs in clinical text as they usually appear in common patterns explicitly and do not need further processing for normalization. For each type of direct TEs, we defined a large number of rules. Taking dates for example, they are calendar dates in three patterns as shown in Figure 3, where “YearDigit”, “MonthNumber” and “DayNumber” are sets of regular expressions for year, month and day numbers respectively (including Chinese numbers such as “零” (zero), “一” (one), “二” (two), etc.), ‘年’, ‘月’ and ‘日/号’ are the units of year, month and day, and “NormYear”, “NormMonth”, “NormDay” are mapping functions from “YearDigit”, “MonthNumber” and “DayNumber” to their standard forms respectively.

**Pattern 1**

- **Extraction rule:** `YearDigit[-一零]+MonthNumber[-一月]+DayNumber[日号]?`
- **Normalization rule:** `NormYear(NormDigit)-NormMonth(NormMonth)-NormDay(NormDay)`
- **Examples:** 2011/6/21, 2011年6月21日 (June 21, 2011), 二零一一年六月二十一日 (June 21, 2011)

**Pattern 2**

- **Extraction rule:** `YearDigit[-一零]+MonthNumber`
- **Normalization rule:** `NormYear(NormDigit)-NormMonth(NormMonth)`
- **Examples:** 2011-6, 2011年6月 (June, 2011), 二零一一年六月 (June, 2011)

**Pattern 3**

- **Extraction rule:** `YearDigit季`
- **Normalization rule:** `NormYear(NormDigit)`

**Figure 3.** Extraction and normalization rules for direct date expressions.
2. Indirect TEs

Compared to direct TEs, indirect TEs are also exact but incomplete or implicit. There are two cases: 1) incomplete temporal expressions missing some parts such as year and month. For example, “6月21日” (“June 21”) is a date expression missing year; 2) implicit temporal expressions representing relative time such as “3天前” (“three days ago”) and “手术前一天” (“the day before surgery”). To exact indirect TEs, we modified rules for direct TEs. To normalize indirect TEs, we need further inference to complete the missing part of them or to calculate the absolute time of them. For example, we need to complete the year information for “6月21日” (“June 21”), and to calculate the absolute time of “3天前” (“three days ago”) and that of “手术前一天” (“the day before surgery”). The key of indirect TE normalization is to find out the baseline time of them which may be a calendar time or the occurring time of some event such as “手术” (“surgery”) in “手术前一天” (“the day before surgery”). In this study, we tried the following three strategies gradually to determine the baseline time of indirect TEs:

1) The baseline time of all indirect TEs in a note is the admission/discharge/create time (for admission notes, discharge summaries and progress notes respectively.) mentioned in the note, which is easy to determine as it usually appears at the beginning of a note.

2) The baseline time of an indirect TE is the nearest TE in front of it in the same sentence or a previous sentence.

3) The baseline time, which is the occurring time of some events (e.g., surgery, operation, transfer, etc.), is determined by the previous two strategies when they first appear.

Figure 4 gave some examples using different strategies for indirect TE normalization, where there was a fragment of an admission note with two direct TEs and five indirect TEs enclosed in square brackets. It is clear that the five indirect TEs were wrongly normalized except the last one when only using strategy 1 mentioned above. When strategy 2 was added, the first three indirect TEs were correctly normalized but the last two TEs not. Nevertheless, strategy 3 corrected the last two TEs normalized by strategy 2 since they were times relative to “手术” (“the operation”) and “入院” (“admission”) with baseline times: 2011-10-18 and 2012-02-03T15:00, respectively. It should be noted that the time of “手术” (“the operation”) is “2011-10-18” not “2011-10-19” since the nearest TE in front of “手术” (“the operation”) that first appears is “18日” (“18th”) not “一天后” (“one day later”).

3. Uncertain TEs

There are two types of uncertain TEs: direct/indirect TEs modified by fuzzy adjectives or adverbs, and fuzzy temporal phrases (e.g., more than, about, in the morning, etc.). For the first type of uncertain TEs, rules for extraction were also modified from direct/indirect TEs by adding fuzzy adjectives or adverbs; and the key of their
normalization lies in a list of modifiers and in which way an uncertain TE and a modifier should be combined together. For the second type of uncertain TEs, we needed to collect a list of fuzzy temporal phrases with modifier attribute. Figure 5 listed some examples of uncertain TEs with modifier value.

<table>
<thead>
<tr>
<th>Direct/indirect TEs with a fuzzy modifier</th>
<th>Fuzzy temporal phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>大约20天</strong> (about 20 days)</td>
<td>目前 (currently)</td>
</tr>
<tr>
<td><strong>早晨8点左右</strong> (about 8 o’clock in the morning)</td>
<td><strong>APPROX</strong></td>
</tr>
<tr>
<td><strong>超过24小时</strong> (more than 24 hours)</td>
<td><strong>APPROX</strong></td>
</tr>
<tr>
<td><strong>4个月前</strong> (more than 4 months)</td>
<td><strong>START</strong></td>
</tr>
<tr>
<td><strong>将近1年</strong> (nearly a year)</td>
<td><strong>YEAR</strong> (the beginning of the year)</td>
</tr>
<tr>
<td><strong>不超过1个月</strong> (less than half a month)</td>
<td><strong>END</strong></td>
</tr>
<tr>
<td><strong>2011年初</strong> (the beginning of 2011)</td>
<td><strong>MID</strong> (the middle of the year)</td>
</tr>
<tr>
<td><strong>明日早晨</strong> (tomorrow morning)</td>
<td><strong>START</strong></td>
</tr>
<tr>
<td><strong>2012年8月</strong> (the middle of 2012)</td>
<td><strong>MIDDLE</strong></td>
</tr>
<tr>
<td><strong>4月30日至6月30日</strong> (the end of April to the evening of 23th)</td>
<td><strong>END</strong></td>
</tr>
</tbody>
</table>

**Figure 5.** Examples of uncertain TEs with modifier value.

**Experiments and evaluation**

We developed our system on the training set, and then evaluated its performance on the independent test set. In order to investigate the effects of the three categories of TEs, we started with the baseline system only considering direct TEs, and then progressively added indirect TEs and uncertain TEs.

The micro-averaged precision, recall, and F-measure under exact/partial-match criteria were used to evaluate system performance for TE extraction, and the micro-averaged accuracy for TE normalization. Under exact-match criterion, an entity is correctly predicted if and only if its text span is exactly the same as that of one entity in the gold standard, while, under partial-match criterion, an entity is correctly predicted if it overlaps with any entity in the gold standard. We developed an evaluation tool based on the official evaluation program provided by the 2012 i2b2 NLP organizers.

**Results**

**Table 3.** Performance of our system on the test set when different categories of TEs were considered (%). “Val” and “Mod” denote the value and modifier attributes.

<table>
<thead>
<tr>
<th>Category</th>
<th>Exact-match</th>
<th></th>
<th>Partial-match</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Span</td>
<td>Type</td>
<td>Val</td>
<td>Mod</td>
</tr>
<tr>
<td>Direct</td>
<td>68.09(79.46/59.57)</td>
<td>59.31</td>
<td>57.88</td>
<td>59.50</td>
</tr>
<tr>
<td>Direct+indirect</td>
<td>82.04(87.00/77.62)</td>
<td>77.28</td>
<td>74.04</td>
<td>77.12</td>
</tr>
<tr>
<td>Direct+indirect+uncertain</td>
<td><strong>93.40(90.83/96.11)</strong></td>
<td><strong>95.72</strong></td>
<td><strong>92.58</strong></td>
<td><strong>95.39</strong></td>
</tr>
</tbody>
</table>

Table 3 showed the performance of our system on the test set when different categories of TEs were considered. The number in columns 2 and 6 were F-measures followed by corresponding precision and recall in parentheses for TE extraction under exact-match and partial-match criteria, while the numbers in other columns were accuracies for identifying corresponding attributes such as numbers in columns 4 and 8 for TE normalization under two criteria. Under exact-match criterion, our system (“direct”) achieved a precision of 79.46%, a recall of 59.57% and an F-measure of 68.09% on TE extraction, and an accuracy of 57.88% on TE normalization when it only considered
direct TEs. When indirect TEs were added (“direct + indirect”), the precision, recall and F-measure of our system on TE extraction were improved to 87.00%, 77.62% and 82.04%, and the accuracy on TE normalization was improved to 74.04%. When further considering uncertain TEs (“direct + indirect + uncertain”), our system achieved highest precision of 90.83%, highest recall of 96.11% and highest F-measure of 93.40% on TE extraction and highest accuracy of 92.58% on TE normalization. These results demonstrated that both indirect and uncertain TEs played an important role in TE extraction and normalization.

The detailed results of the best system for each type of TE are shown in Table 4. On TE extraction, F-measures ranged from 88.45% to 97.92% under exact-match criterion. Among four types of TEs, our system performed best for times and worse for durations. On TE normalization, accuracies ranged from 87.69% to 96.48% under exact-match criterion, and our system performed best for times and worse for frequencies.

<table>
<thead>
<tr>
<th>Type</th>
<th>Exact-match</th>
<th>Partial-match</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Span</td>
<td>Type</td>
</tr>
<tr>
<td>Overall</td>
<td>93.40(90.83/96.11)</td>
<td>95.72</td>
</tr>
<tr>
<td>Date</td>
<td>93.65(90.79/96.69)</td>
<td>96.69</td>
</tr>
<tr>
<td>Time</td>
<td>97.92(97.98/97.87)</td>
<td>97.87</td>
</tr>
<tr>
<td>Duration</td>
<td>88.45(82.86/94.85)</td>
<td>94.85</td>
</tr>
<tr>
<td>Frequency</td>
<td>93.30(92.97/93.62)</td>
<td>93.62</td>
</tr>
</tbody>
</table>

Discussion

In this paper, we comprehensively investigated the rule-based methods for TE extraction and normalization in Chinese clinical notes. We manually created an annotated dataset of 1,778 clinical notes in Chinese for system development and test. Experiments on the dataset showed that our system achieved the highest F-measure of 93.40% on TE extraction, and the highest accuracy of 92.58% on TE normalization under exact-match criterion. We published our system online (http://icrc.hitsz.edu.cn:8096/CMedTEX), which is the first public available TE extraction and normalization system for Chinese clinical notes and will be useful for future Chinese clinical NLP studies. The source code of our system was available for application at: http://icrc.hitsz.edu.cn/Article/show/147.html.

It is clear that our system performance is progressively improved when gradually adding specific rules for direct, indirect and uncertain TEs as shown in Table 3. The main reason lies in that when the rules for direct TEs are deployed to indirect and uncertain TEs, parts of indirect and uncertain TEs will be recognized as direct TEs, which will cause some errors. For example, when applying rules for direct TEs to “2011 年 6 月 21 日下午” (“afternoon of June 21, 2011”), an indirect TE, and “3 天以前” (“3 days ago”), an uncertain TE, we will only obtain two direct TEs: a date of “2011 年 6 月 21 日” (“June 21, 2011”) and a duration of “3 天” (“3 days”), which are two errors. Therefore, we designed specific rules for indirect and uncertain TEs, respectively, which are derived from the rules for direct TEs, to correct these errors.

Compared to TE extraction and normalization systems in other domains (e.g., newswire) or other languages (e.g., English), our system is promising. For example, the best system on the Chinese corpus of TempEval-2 in newswire domain (i.e., HeidelTime) was rule-based and achieved an F-measure of 89.30% on TE extraction and an accuracy of 87% on TE normalization under exact-match criterion. In the 2012 i2b2 challenge about temporal information extraction for English clinical text, the highest F-measure and accuracy on TE extraction and normalization were 91.44% and 73.13% respectively. Our system showed better performance than the best systems of those two challenges even though the comparison between them is not very fair as they were not evaluated under the same conditions.

In order to investigate the effect of domain difference on TE extraction and normalization systems, we compared HeidelTime, a state of the art TE extraction and normalization system for Chinese newswire text, with our system on the test set. Under exact-match criterion, the F-measure and accuracy on TE extraction and normalization of HeidelTime were only 38.35% and 8.60%, much lower than those of our system, indicating that TEs in clinical text are much different from TEs in other domains (especially newswire domain). Therefore, it is necessary to design...
new rules and systems to extract and normalize TEs in clinical text. The reasons why HeidelTime performed so bad for Chinese clinical notes may lie in the following two aspects: 1) there are a large number of proper TEs in clinical text such as qid (“一天四次”, four times per day), which cannot be recognized by HeidelTime; and 2) many indirect TEs in clinical text are relative to clinical events such as operation and transfer, which are not considered by HeidelTime.

Furthermore, we analyzed the errors caused by our system, and found that most of the errors were related to word ambiguities. For example, “9-10” can be either a date “9月10号” (“September 10”) or a duration “9至10” (“from 9 to 11”), “9.10” can be either a date “9月10号” (“September 10”) or a number, “10个月” (“10 months”) can be either a duration or a child’s age, “夜间” (“at night”) is usually a duration, but not a separate TE when it is a part of problem “夜间盗汗” (“night sweats”), and “一周” (“one week”) is usually a duration, but not a TE when it appears in problem “脐带绕颈一周” (“a nuchal cord”) with another meaning “one round” in Chinese. Among these errors, the first three are common problems in both English and Chinese clinical text and the last two are special problems in Chinese clinical text. To tackle these problems, we may need to take other information, such as syntax and clinical entities, into account. However, it is not easy to obtain these information as the studies on Chinese clinical NLP has just began.

In the future, we plan to focus on the following two aspects for further improvement: 1) consider more information, such as syntactic and semantic information, to fix current errors; and 2) attempt machine learning-based methods to enhance our system.

Conclusion

In this study, we developed a rule-based time expression extraction and normalization system for Chinese clinical notes. The time expressions were summarized into three different categories: direct, indirect and uncertain expressions, which facilitated the construction of rules more easily and clearly. Our system achieved a span F-measure of 93.40% and a value accuracy of 92.58% under “exact-match” criterion. The online system is publicly available and its source code is released for application.

Acknowledgements

This paper is supported in part by grants: National 863 Program of China (2015AA015405), NSFCs (National Natural Science Foundations of China) (61573118, 61402128, 61473101, 61472428, 61173075 and 61272383), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20140508161040764, JCYJ20140417172417105, JCYJ20140627163809422 and JSGG20151015161015297), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052) and Program from the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (93K172016K12).

References


Interpretable Topic Features for Post-ICU Mortality Prediction

Yen-Fu Luo, MS, Anna Rumshisky, PhD
University of Massachusetts Lowell, Lowell, MA

Abstract

Electronic health records provide valuable resources for understanding the correlation between various diseases and mortality. The analysis of post-discharge mortality is critical for healthcare professionals to follow up potential causes of death after a patient is discharged from the hospital and give prompt treatment. Moreover, it may reduce the cost derived from readmissions and improve the quality of healthcare.

Our work focused on post-discharge ICU mortality prediction. In addition to features derived from physiological measurements, we incorporated ICD-9-CM hierarchy into Bayesian topic model learning and extracted topic features from medical notes. We achieved highest AUCs of 0.835 and 0.829 for 30-day and 6-month post-discharge mortality prediction using baseline and topic proportions derived from Labeled-LDA. Moreover, our work emphasized the interpretability of topic features derived from topic model which may facilitates the understanding and investigation of the complexity between mortality and diseases.

1. Introduction

Post-discharge management is one of the important aspects in current healthcare system. For high-risk patients, and especially for the intensive care unit (ICU) patients, it is critical to understand and prevent possible complications and problems which may lead to a patient’s death after being discharged from the hospital. The present work focused on mortality prediction of high-risk ICU patients. In our patient cohort, the post-discharge mortality for 30-day and 6-month are 3.4% and 9.5% respectively. There has been a lot of recent interest in mortality prediction in general and post-ICU mortality prediction in particular\(^1\)\(^-\)\(^3\). However, many of the state-of-the-art methods use “black box” predictive models which can not provide any explanation for practitioners as to why a particular patient may be at risk after discharge. In this paper, our goal is two-fold: develop novel methods that can both accurately predict mortality and at the same time create a transparent predictive model that can be easily understood and therefore actionable by the providers.

SAPS-II\(^4\), APACHE-II\(^7\), and SOFA\(^6\) scores are commonly used in ICU mortality prediction\(^1\)\(^-\)\(^2\), \(^7\)\(^-\)\(^10\). In addition to structured data and derived severity scores, we build a mortality prediction model that incorporates features derived from unstructured medical notes. We use Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II)\(^11\) database. The narrative provider notes from MIMIC II give detailed descriptions of symptoms, diagnosis, surgery, medicine, and treatments. It is highly informative but in the form of free-text. Our goal is to capture clinically relevant information and patterns identified and summarized by healthcare providers in order to leverage them in transparent prediction.

There have been recent attempts to use Bayesian topic modeling techniques to improve mortality prediction using narrative notes\(^1\)\(^-\)\(^5\), \(^7\). In topic modeling, each document is represented as a probability distribution over a set of topics and each topic is modeled as a probability distribution over a set of words. Although topic-based features have been used in literature to improve outcome prediction, the topics themselves are flat word collections that need to be examined by domain expert in order to assign a clinical interpretation.

Although the derived topics show some degree of interpretability\(^12\), human annotators are prone to assigning meaning to topics or word clusters even in cases when such word collections are not coherent. In this work, we propose a method to automatically define interpretable topics. To make a topic itself interpretable and clearly definable based on domain knowledge, we used the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes as topics namely labels in Labeled-LDA to guide topic model learning and extract understandable topic feature representations from medical notes. In addition, we examined the feasibility of using topic features derived from Labeled-LDA for post-discharge mortality prediction.

2. Related Work

One of the common approaches to using narrative notes for clinical outcome prediction is to extract clinically-relevant concept and relations using information extraction techniques and use them as features in predictive models. A number of medical concept extraction systems, including rule-based systems (MetaMap\(^13\), MedLEE\(^14\)
cTAKES\textsuperscript{15}, etc.) and machine learning-based systems (Clien\textsuperscript{16}, RapTAT\textsuperscript{17}) have been used for this task\textsuperscript{15-23}. However, such systems require a substantial amount of human labor in order to produce accurate results, either for rule construction and keyword selection, or for text annotation and feature engineering required for the supervised machine learning. As a result, shifting between different types of clinical notes, or between different institutions, requires a substantial overhead in order to achieve domain adaptation. This is additionally compounded by continuous changes in medical terminology, introduction of new medication brands, and so on.

There has been a number of recent attempts to bypass this problem by using unsupervised methods that rely on topic modeling to extract topic features from clinical narrative text in order to improve the prediction of in-hospital and post-discharge mortality for ICU patients. Ghassemi et al.\textsuperscript{1} reported AUCs of 0.754 and 0.781 for 30-day and 6-month post-discharge mortality prediction. Ghassemi et al.\textsuperscript{2} also reported an AUC of 0.818 for 30-day post-discharge mortality prediction using retrospective topic + derived features model. Lehman et al.\textsuperscript{7} combined medical concepts extracted from medical notes with topic model for ICU in-hospital mortality prediction. Jo et al.\textsuperscript{3} used state-transition topic model for incorporating temporal information and reached an AUC of 0.792 for 6-month post-discharge mortality prediction.

Ideally, predictive models for mortality should be customized for different patient groups, based on primary diagnosis and other patient characteristics. Nori et al.\textsuperscript{24} combined the hierarchy of ICD-10 into mortality prediction and divided the general prediction model into multi-task/multi-disease learning problem. Makar et al.\textsuperscript{25} also incorporated ICD-9-CM codes for short-term mortality prediction of elderly patients.

Topic models such as the Latent Dirichlet Allocation (LDA)\textsuperscript{26} and Hierarchical Dirichlet Process (HDP)\textsuperscript{27} are widely used to explore coherent topics within large text corpora. HDP is a nonparametric Bayesian approach which does not require specifying the desired number of topics. Arnold et al.\textsuperscript{12} showed the interpretability of topic model from a physician’s perspective. Although their conclusions support using topic features in a prediction task, identifying high quality topics may also require a labor-intensive topic evaluation by domain experts, in order to determine the optimal parameter settings (i.e., the number of topics in LDA or the concentration hyperparameter in HDP). Incorporating domain knowledge into the topic learning in the way implemented in the present work helps to address both the problem of customizing predictive models for different patient categories and to improve topic interpretability.

3. Methods

In the present work, we propose to incorporate domain knowledge into topic learning using Labeled-LDA\textsuperscript{28} with ICD-9-CM codes as labels. Labels are equivalent to topics or ICD-9-CM codes in our setting, and each document may be assigned multiple labels. In LDA, all documents contribute to all topics in the learning of the topic model. In Labeled-LDA, a subset of documents with the corresponding label is used to infer word distributions for a topic. The benefit of using ICD-9-CM codes as labels in Labeled-LDA is two-fold. First, the clinical notes from a given patient’s record contribute only to a subset of topics corresponding to the ICD-9-CM code assignments for that patient. Second, topic interpretability is achieved through a combination of the ICD-9-CM code definition and the top words for a given topic. At the training stage, we incorporate ICD-9-CM codes to guide the Labeled-LDA model learning. However, since ICD-9-CM codes are not available at the time of discharge, they can not be included as features in the predictive models directly. Using Labeled-LDA approach allows us to bypass this issue, since at the prediction stage, inferring topic proportions corresponding to different ICD-9-CM codes does not require one to have the ICD-9-CM codes available.

3.1 Patient Selection

We used MIMIC II database. The database contains physiological signals, vital signs, medical notes, and other structured data from several ICUs, including medical, surgical, coronary care, and neonatal. This data was collected between 2001 and 2008 at Boston’s Beth Israel Deaconess Medical Center (BIDMC). The database contains over 25,000 patients including around 20,000 adults and 5,000 neonates.

Since the factors related to mortality differ substantially for neonates, only patients in the adult age group were selected. The patients without discharge summary were excluded from the cohort, since discharge summary is essential for building the prediction model for post-discharge mortality. We also excluded patients without the first SAPS II score. All available clinical notes, including nursing, physician, radiology, and the discharge summary of the patient’s first hospital stay were collected. We identified and removed 11% duplicate medical notes. The resulting cohort consisted of 18,412 patients with 400,494 notes. The patient data was randomly split into 80% for training set and 20% for testing set.
3.2 Preprocessing and Tokenization

Each note was processed by using the SPECIALIST Lexicon LRABR table to preserve medical abbreviations and acronyms; this was followed by whitespace-based tokenization, and the removal of stopwords from the Onix stopword list. Term frequency was generated by aggregating word frequency for each patient. The top 500 most informative words were selected for each patient, based on the TF-IDF score29 for each patient included in the training data. This resulted in the overall vocabulary containing 151,772 words.

3.3 LDA and Labeled-LDA

Knowledge Based Topic Models (KBTM) were developed to guide topic model learning by incorporating domain knowledge. Andrzejewski et al.30-31 demonstrated the use of Dirichlet forest priors and first-order logic in order to create must-links and cannot-links between words which encode domain knowledge during model learning. This solution requires domain experts to encode knowledge used to create constraints. We propose using Labeled-LDA as an alternative. Labeled-LDA was designed to analyze the text of the web pages that may be annotated by users in a community portal. Each page may have multiple labels associated with the topics of the page (such as arts, politics, physics, religion, alaska, etc.), assigned by the readers. In Labeled-LDA, word distribution of a topic is inferred based on a subset of the corpus with the corresponding label. Therefore, the inferred top keywords of a topic are associated with the subject of the label.

We adapted this model to the task of transparent outcome prediction, using the ICD-9 diagnostic codes assigned to each patient as labels, with the two-fold goal of guiding the topic learning and improving interpretability of the resulting topics. ICD-9-CM is based on the World Health Organization’s Ninth Revision, International Classification of Diseases (ICD-9). Until October 2015, diagnoses and procedures associated with hospital utilization in the United States were recorded using official ICD-9-CM codes. Based on multiple procedures and treatments during a patient’s hospital stay, multiple ICD-9-CM codes are assigned by trained healthcare professionals. Although there is some disagreement on the viability of using ICD-9 codes in predictive tasks due to the diagnostic codes being assigned exclusively for billing purposes, they do provide an expert-generated authoritative source of annotation for each record, which can be reasonably assumed to represent high-level domain knowledge. The topics obtained by applying Labeled-LDA to medical notes of ICU patients with ICD-9-CM codes as labels may be interpreted as providing a description of sorts to the corresponding code, which can be easily verified against the ICD-9 code definition. Under this interpretation, the topic proportions also represent the extent to which a particular diagnosis or procedure is associated with the given hospital stay.

Rather than using the raw ICD-9-CM codes, we used the ICD-9-CM hierarchy with 180 upper-level codes in order to reduce the sparsity of ICD-9-CM code assignments. Comparing to the LDA model that uses all documents to infer topic proportions and word distributions, the Labeled-LDA model only uses a subset of documents for each topic inference. Since training a topic model requires sufficient data to produce coherent topics, we only considered labels with minimum frequency 50, 100, 200, and 400. The resulting number of labels are 111, 94, 79, and 59 respectively. Following Ghassemi et al1-2, we used 50 topics for training a regular LDA model. We sampled topic proportions for each patient in the training data after using 2,500 iterations during model learning, and the resulting model was used to obtain topic proportions for each patient in the test data. We used default hyperparameter settings in both LDA and Labeled-LDA models.

3.4 Mortality Prediction

We retrieved age, gender, SAPS-II scores, Elixhauser Comorbidity Index32, the text of the medical notes, and the ICD-9-CM code assignments for each patient’s first hospital stay recorded in the MIMIC II database. The topic model was used to infer topic proportions for all medical notes in a patient’s record at the time of discharge. This information, together with the obtained topic proportions, were used as real-valued features in a predictive model. A Support Vector Machine (SVM) model33 with radial basis function (RBF) kernel was trained and used to predict 30-day and 6-month post-discharge mortality.

For 6-month post-discharge mortality prediction, we used three feature settings: (1) baseline features included age, gender, SAPS-II score at admission, minimum SAPS II score, maximum SAPS II score, and 30 Elixhauser Comorbidities, (2) baseline features and 50 topic proportions derived from regular LDA, and (3) baseline features and topic proportions derived from Labeled-LDA. For the 30-day prediction model, we excluded minimum SAPS-II score, maximum SAPS-II score, and ICD-9-CM derived 30 Elixhauser Comorbidities from the baseline features.
The reason is that the assignment of ICD-9-CM codes is usually finalized within 2 weeks after a patient is discharged from the hospital, and therefore is not available at the time the prediction needs to be made.

In our patient cohort, 3.4% and 9.5% of the patients died within 30-day and 6-month post-discharge respectively. Because of the highly imbalanced data, we subsampled negative class to generate dataset with 20% positive and 80% negative class for training. In addition, we penalized misclassification of positive class by assigning higher class weight in SVM. The optimal cost and gamma parameters were determined in 5-fold cross-validation over the training data against ROC-AUC.

4. Results

4.1 Topic Interpretability

Baseline + Labeled-LDA with 111 labels achieved highest AUC in both 30-day and 6-month mortality prediction. To illustrate the topics derived with this Labeled-LDA model, we show 10 most and 10 least frequent ICD-9-CM codes in Table 1, along with their definitions, and top 20 words. The results suggest the consistency between ICD-9-CM code’s definition and the corresponding keywords. For example, the top words for the “hypertensive disease” topic include ‘chest’, ‘cabg’, ‘artery’, ‘coronary’, etc. Another example topic, labeled “complications occurring mainly in the course of labor and delivery” is associated with the words ‘uterine’, ‘bleeding’, ‘vaginal’, ‘delivery’, ‘abd’, and ‘hct’.

Table 1. List of top 20 words learned from Labeled-LDA and its corresponding ICD-9-CM definition. Top 10 and bottom 10 entries are most and least frequent ICD-9-CM code in our dataset. Frequency of 111 ICD-9-CM codes are listed in Table 3.

<table>
<thead>
<tr>
<th>ICD-9-CM</th>
<th>Definition (above) / Keywords (below)</th>
</tr>
</thead>
<tbody>
<tr>
<td>401-405</td>
<td>Hypertensive disease</td>
</tr>
<tr>
<td></td>
<td>tablet chest left mg po sig pt daily reason cabg artery sp refills disp namepattern clip pain date day coronary</td>
</tr>
<tr>
<td>420-429</td>
<td>Other forms of heart disease</td>
</tr>
<tr>
<td></td>
<td>pt mg patient hr chest resp left lasix gi po stable pain gu neuro gtt bp bs day cv plan</td>
</tr>
<tr>
<td>270-279</td>
<td>Other metabolic and immunity disorders</td>
</tr>
<tr>
<td></td>
<td>patient mg pt chest day left artery pain po stable coronary cabg status discharge history date post namepattern clip examination</td>
</tr>
<tr>
<td>410-414</td>
<td>Ischemic heart disease</td>
</tr>
<tr>
<td></td>
<td>mg pt patient cath tablet pain left cardiac chest po hospital hr artery ccu discharge coronary history normal namepattern daily</td>
</tr>
<tr>
<td>249-259</td>
<td>Diseases of other endocrine glands</td>
</tr>
<tr>
<td></td>
<td>pt patient mg insulin day blood hr po pain discharge units bs diabetes history namepattern hospital gtt admission pm doctor</td>
</tr>
<tr>
<td>510-519</td>
<td>Other diseases of respiratory system</td>
</tr>
<tr>
<td></td>
<td>pt hr resp vent remains cc secretions thick care tube plan bs neuro trach cont mg gi noted yellow abg</td>
</tr>
<tr>
<td>996-999</td>
<td>Complications of surgical and medical care, not elsewhere classified</td>
</tr>
<tr>
<td></td>
<td>pt tube resp left hr chest plan neuro remains cc vent bs cont reason noted clip abd gi sp care</td>
</tr>
<tr>
<td>280-285</td>
<td>Anemia</td>
</tr>
<tr>
<td></td>
<td>pt tablet mg blood po hct sig daily discharge pm doctor namepattern md patient pain day history gi admission hospital</td>
</tr>
<tr>
<td>780-789</td>
<td>Symptoms</td>
</tr>
<tr>
<td></td>
<td>patient pt contrast ct head left clip seizure reason normal pm date mri mg evidence hospital report history examination noted</td>
</tr>
<tr>
<td>580-589</td>
<td>Nephritis, nephrotic syndrome, and nephrosis</td>
</tr>
<tr>
<td></td>
<td>renal clip reason left chest failure line final catheter radiology report examination date medical underlying patient pleural condition dialysis hd</td>
</tr>
<tr>
<td>317-319</td>
<td>Mental retardation</td>
</tr>
<tr>
<td></td>
<td>pt tube noted chest cc resp patient retardation thick secretions care cont plan trach abd hr ct neuro telemetry coarse</td>
</tr>
<tr>
<td>E910-915</td>
<td>Accidents caused by submersion, suffocation, and foreign bodies</td>
</tr>
<tr>
<td></td>
<td>pt patient esophageal food namepattern care perforation pain impaction oral secretions aspiration time esophagus white wife hospital discharge doctor intubated</td>
</tr>
</tbody>
</table>
Table 1. (Continued)

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>950-957</td>
<td>Injury to nerves and spinal cord</td>
</tr>
<tr>
<td></td>
<td>pt resp trach family care pain neuro injury plan vent intact hr gi skin thick secretions movement noted cord yellow</td>
</tr>
<tr>
<td>338-338</td>
<td>Pain</td>
</tr>
<tr>
<td></td>
<td>pain pt mg tablet po sig doctor patient daily md blood ml iv namepattern discharge prn disp refills hr esophageal</td>
</tr>
<tr>
<td>905-909</td>
<td>Late effects of injuries, poisonings, toxic effects, and other external causes</td>
</tr>
<tr>
<td></td>
<td>pt noted care pain intact wound skin cont vac family patient yellow vent drainage plan changed secretions abd resp remains</td>
</tr>
<tr>
<td>910-919</td>
<td>Superficial injury</td>
</tr>
<tr>
<td></td>
<td>signal ml thoracic level ativan images foraminal spine fentanyl stenosis pm moderate pressure sbp seizures jump mild ligamentous abrasions ointment</td>
</tr>
<tr>
<td>E820-825</td>
<td>Motor vehicle non-traffic accidents</td>
</tr>
<tr>
<td></td>
<td>pt trauma contrast family vehicle motor neuro mva head hr remains support intact skin ct trach mvc vent sp mri</td>
</tr>
<tr>
<td>890-897</td>
<td>Open wound of lower limb</td>
</tr>
<tr>
<td></td>
<td>pt resp skin care support thick intact wound plan family secretions remains peep tube vent drainage hr stable cont bs</td>
</tr>
<tr>
<td>V20-29</td>
<td>Persons encountering health services in Circumstances related to Reproduction and development</td>
</tr>
<tr>
<td></td>
<td>pt hr drainage abd continue continues support vent hct fluid family ativan mg husband cont skin cv resp white line</td>
</tr>
<tr>
<td>660-669</td>
<td>Complications occurring mainly in the course of labor and delivery</td>
</tr>
<tr>
<td></td>
<td>pt patient blood pm uterine bleeding clip hct post reason vaginal date history abd artery units namepattern delivery discharge sp</td>
</tr>
</tbody>
</table>

4.2 Mortality Prediction

Table 2 shows mortality prediction results, with AUC, sensitivity, and specificity shown for baseline features, baseline + LDA topics, and baseline + Labeled-LDA topics with four label settings. The model using baseline + topic features from Labeled-LDA with 111 labels achieved an AUC of 0.835 for 30-day post-discharge mortality prediction. For 6-month post-discharge mortality prediction, baseline + Labeled-LDA with 111 and 94 labels performed closely with AUCs of 0.829. While both topic model derived features outperform the baseline in both 30-day and 6-month prediction model, baseline + LDA topics achieves somewhat higher AUCs than baseline + Labeled-LDA topics.

Table 2. Results of 30-day and 6-month mortality prediction.

<table>
<thead>
<tr>
<th>Post-discharge Timeframe</th>
<th>Prediction Model</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-day</td>
<td>baseline</td>
<td>0.736</td>
<td>75.000</td>
<td>56.063</td>
</tr>
<tr>
<td></td>
<td>baseline + LDA with 50 topics</td>
<td>0.860</td>
<td>86.607</td>
<td>70.569</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 111 labels</td>
<td>0.835</td>
<td>85.714</td>
<td>63.204</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 94 labels</td>
<td>0.834</td>
<td>86.607</td>
<td>63.652</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 79 labels</td>
<td>0.832</td>
<td>86.607</td>
<td>63.596</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 59 labels</td>
<td>0.831</td>
<td>89.286</td>
<td>59.563</td>
</tr>
<tr>
<td>6-month</td>
<td>baseline</td>
<td>0.776</td>
<td>71.831</td>
<td>70.343</td>
</tr>
<tr>
<td></td>
<td>baseline + LDA with 50 topics</td>
<td>0.842</td>
<td>78.873</td>
<td>75.090</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 111 labels</td>
<td>0.829</td>
<td>78.873</td>
<td>73.137</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 94 labels</td>
<td>0.829</td>
<td>78.592</td>
<td>71.545</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 79 labels</td>
<td>0.827</td>
<td>78.873</td>
<td>72.176</td>
</tr>
<tr>
<td></td>
<td>baseline + Labeled-LDA with 59 labels</td>
<td>0.826</td>
<td>78.873</td>
<td>71.154</td>
</tr>
</tbody>
</table>
4.3 Topic Mortality

We applied probability of mortality defined by Marlin et al. for each topic to investigate correlation between topics and mortality. Table 3 depicts ICD-9-CM codes with corresponding probability of mortality for 30-day and 6-month post-discharge periods. The results suggested “viral diseases accompanied by exanthem” (050-059), “dislocation” (830-839), and malignant neoplasm of “other and unspecified sites”, “respiratory and intrathoracic organs” (190-199 and 160-165), and “other diseases of skin and subcutaneous tissue” (700-709) are the potentially important causes of death for 30-day post-discharge. For 6-month post-discharge mortality, malignant neoplasm of “other and unspecified sites”, “respiratory and intrathoracic organs” (190-199, 160-165, and 200-208), and “other diseases of skin and subcutaneous tissue” (700-709) are the potentially important causes of death. On the other hand, one can see that “open wound of limb” (880-887 and 890-897), “superficial injury” (910-919), “complications of labor and delivery” (660-669), “complications mainly related to pregnancy” (640-649), “injury to blood vessels” (900-904), “homicide and injury purposely inflicted by other persons” (E960-969), and “suicide and self-inflicted injury” (E950-959) were ranked high in both 30-day and 6-month post-discharge survival.

Table 3. The probability of mortality for 111 topics for 30-day and 6-month; the top potential causes of death are highlighted in bold; frequency of ICD-9-CM codes

<table>
<thead>
<tr>
<th>ICD-9-CM</th>
<th>Definition</th>
<th>30-day</th>
<th>6-month</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>001-009</td>
<td>Intestinal infectious diseases</td>
<td>0.0607</td>
<td>0.1752</td>
<td>371</td>
</tr>
<tr>
<td>030-041</td>
<td>Other bacterial diseases</td>
<td>0.0569</td>
<td>0.1404</td>
<td>1875</td>
</tr>
<tr>
<td>042-044</td>
<td>Human immunodeficiency virus (HIV) infection</td>
<td>0.0336</td>
<td>0.1203</td>
<td>149</td>
</tr>
<tr>
<td>050-059</td>
<td>Viral diseases accompanied by exanthem</td>
<td>0.1644</td>
<td>0.1693</td>
<td>85</td>
</tr>
<tr>
<td>070-079</td>
<td>Other diseases due to viruses and chlamydiae</td>
<td>0.0467</td>
<td>0.1365</td>
<td>593</td>
</tr>
<tr>
<td>110-118</td>
<td>Mycoses</td>
<td>0.0707</td>
<td>0.1483</td>
<td>410</td>
</tr>
<tr>
<td>130-136</td>
<td>Other infectious and parasitic diseases</td>
<td>0.0367</td>
<td>0.0664</td>
<td>87</td>
</tr>
<tr>
<td>150-159</td>
<td>Malignant neoplasm of digestive organs and peritoneum</td>
<td>0.0271</td>
<td>0.1146</td>
<td>415</td>
</tr>
<tr>
<td>160-165</td>
<td>Malignant neoplasm of respiratory and intrathoracic organs</td>
<td>0.1190</td>
<td>0.2723</td>
<td>289</td>
</tr>
<tr>
<td>170-175</td>
<td>Malignant neoplasm of bone, connective tissue, skin, and breast</td>
<td>0.0072</td>
<td>0.1139</td>
<td>79</td>
</tr>
<tr>
<td>179-189</td>
<td>Malignant neoplasm of genitourinary organs</td>
<td>0.0355</td>
<td>0.0826</td>
<td>222</td>
</tr>
<tr>
<td>190-199</td>
<td>Malignant neoplasm of other and unspecified sites</td>
<td>0.1134</td>
<td>0.3519</td>
<td>768</td>
</tr>
<tr>
<td>200-208</td>
<td>Malignant neoplasm of lymphatic and hematopoietic tissue</td>
<td>0.0764</td>
<td>0.2443</td>
<td>292</td>
</tr>
<tr>
<td>210-229</td>
<td>Benign neoplasms</td>
<td>0.0095</td>
<td>0.0332</td>
<td>331</td>
</tr>
<tr>
<td>235-238</td>
<td>Neoplasms of uncertain behavior</td>
<td>0.0949</td>
<td>0.2122</td>
<td>152</td>
</tr>
<tr>
<td>240-246</td>
<td>Disorders of thyroid gland</td>
<td>0.0267</td>
<td>0.0939</td>
<td>1264</td>
</tr>
<tr>
<td>249-259</td>
<td>Diseases of other endocrine glands</td>
<td>0.0270</td>
<td>0.0840</td>
<td>4022</td>
</tr>
<tr>
<td>260-269</td>
<td>Nutritional deficiencies</td>
<td>0.0726</td>
<td>0.2033</td>
<td>369</td>
</tr>
<tr>
<td>270-279</td>
<td>Other metabolic and immunity disorders</td>
<td>0.0094</td>
<td>0.0356</td>
<td>6912</td>
</tr>
<tr>
<td>280-285</td>
<td>Anemia</td>
<td>0.0382</td>
<td>0.1156</td>
<td>3731</td>
</tr>
<tr>
<td>286-287</td>
<td>Coagulation/hemorrhagic</td>
<td>0.0843</td>
<td>0.1648</td>
<td>1170</td>
</tr>
<tr>
<td>288-289</td>
<td>Other</td>
<td>0.0491</td>
<td>0.1894</td>
<td>289</td>
</tr>
<tr>
<td>290-294</td>
<td>Organic psychotic conditions</td>
<td>0.0674</td>
<td>0.1359</td>
<td>1219</td>
</tr>
<tr>
<td>295-299</td>
<td>Other Disorders</td>
<td>0.0245</td>
<td>0.0450</td>
<td>494</td>
</tr>
<tr>
<td>300</td>
<td>Neurotic disorders</td>
<td>0.0021</td>
<td>0.0332</td>
<td>450</td>
</tr>
<tr>
<td>303-305</td>
<td>Psychoactive substance</td>
<td>0.0079</td>
<td>0.0295</td>
<td>1773</td>
</tr>
<tr>
<td>306-311</td>
<td>Other (primarily adult onset)</td>
<td>0.0182</td>
<td>0.0667</td>
<td>707</td>
</tr>
<tr>
<td>317-319</td>
<td>Mental retardation</td>
<td>0.0110</td>
<td>0.0259</td>
<td>71</td>
</tr>
<tr>
<td>320-327</td>
<td>Inflammatory diseases of the central nervous system</td>
<td>0.0277</td>
<td>0.0705</td>
<td>377</td>
</tr>
<tr>
<td>330-337</td>
<td>Hereditary and degenerative diseases of the central nervous system</td>
<td>0.0319</td>
<td>0.1157</td>
<td>600</td>
</tr>
<tr>
<td>338-338</td>
<td>Pain</td>
<td>0.0219</td>
<td>0.0400</td>
<td>60</td>
</tr>
<tr>
<td>340-349</td>
<td>Other disorders of the central nervous system</td>
<td>0.0208</td>
<td>0.0768</td>
<td>938</td>
</tr>
<tr>
<td>350-359</td>
<td>Disorders of the peripheral nervous system</td>
<td>0.0065</td>
<td>0.0620</td>
<td>611</td>
</tr>
<tr>
<td>360-379</td>
<td>Disorders of the eye and adnexa</td>
<td>0.0331</td>
<td>0.0848</td>
<td>617</td>
</tr>
<tr>
<td>380-389</td>
<td>Diseases of the ear and mastoid process</td>
<td>0.0501</td>
<td>0.1752</td>
<td>119</td>
</tr>
<tr>
<td>393-398</td>
<td>Chronic rheumatic heart disease</td>
<td>0.0156</td>
<td>0.0499</td>
<td>585</td>
</tr>
<tr>
<td>401-405</td>
<td>Hypertensive disease</td>
<td>0.0058</td>
<td>0.0200</td>
<td>7452</td>
</tr>
<tr>
<td>410-414</td>
<td>Ischemic heart disease</td>
<td>0.0184</td>
<td>0.0502</td>
<td>5416</td>
</tr>
<tr>
<td>Code</td>
<td>Description</td>
<td>Year 1</td>
<td>Year 2</td>
<td>Total</td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>415-417</td>
<td>Diseases of pulmonary circulation</td>
<td>0.0609</td>
<td>0.1183</td>
<td>763</td>
</tr>
<tr>
<td>420-429</td>
<td>Other forms of heart disease</td>
<td>0.0394</td>
<td>0.1068</td>
<td>7165</td>
</tr>
<tr>
<td>430-438</td>
<td>Cerebrovascular disease</td>
<td>0.0358</td>
<td>0.0745</td>
<td>1562</td>
</tr>
<tr>
<td>440-448</td>
<td>Diseases of arteries, arterioles, and capillaries</td>
<td>0.0175</td>
<td>0.0620</td>
<td>1521</td>
</tr>
<tr>
<td>451-459</td>
<td>Diseases of veins and lymphatics, and other diseases of circulatory system</td>
<td>0.0624</td>
<td>0.1487</td>
<td>1880</td>
</tr>
<tr>
<td>460-466</td>
<td>Acute respiratory infections</td>
<td>0.0394</td>
<td>0.0570</td>
<td>121</td>
</tr>
<tr>
<td>470-478</td>
<td>Other diseases of the upper respiratory tract</td>
<td>0.0250</td>
<td>0.0514</td>
<td>215</td>
</tr>
<tr>
<td>480-488</td>
<td>Pneumonia and influenza</td>
<td>0.0439</td>
<td>0.1523</td>
<td>1889</td>
</tr>
<tr>
<td>490-496</td>
<td>Chronic obstructive pulmonary disease and allied conditions</td>
<td>0.0537</td>
<td>0.1452</td>
<td>2452</td>
</tr>
<tr>
<td>500-508</td>
<td>Pneumococcosis and other lung diseases due to external agents</td>
<td>0.0744</td>
<td>0.1845</td>
<td>1059</td>
</tr>
<tr>
<td>510-519</td>
<td>Other diseases of respiratory system</td>
<td>0.0725</td>
<td>0.1877</td>
<td>3904</td>
</tr>
<tr>
<td>520-529</td>
<td>Diseases of oral cavity, salivary glands, and jaws</td>
<td>0.0147</td>
<td>0.0504</td>
<td>152</td>
</tr>
<tr>
<td>530-537</td>
<td>Diseases of esophagus, stomach, and duodenum</td>
<td>0.0266</td>
<td>0.0747</td>
<td>2200</td>
</tr>
<tr>
<td>550-553</td>
<td>Hernia of abdominal cavity</td>
<td>0.0240</td>
<td>0.0486</td>
<td>314</td>
</tr>
<tr>
<td>555-558</td>
<td>Noninfectious enteritis and colitis</td>
<td>0.0221</td>
<td>0.1213</td>
<td>364</td>
</tr>
<tr>
<td>560-569</td>
<td>Other diseases of intestines and peritoneum</td>
<td>0.0475</td>
<td>0.1165</td>
<td>1353</td>
</tr>
<tr>
<td>570-579</td>
<td>Other diseases of digestive system</td>
<td>0.0567</td>
<td>0.1277</td>
<td>2025</td>
</tr>
<tr>
<td>580-589</td>
<td>Nephritis, nephrotic syndrome, and nephrosis</td>
<td>0.0771</td>
<td>0.1749</td>
<td>2824</td>
</tr>
<tr>
<td>590-599</td>
<td>Other diseases of urinary system</td>
<td>0.0716</td>
<td>0.1831</td>
<td>2401</td>
</tr>
<tr>
<td>600-608</td>
<td>Diseases of male genital organs</td>
<td>0.0225</td>
<td>0.0925</td>
<td>434</td>
</tr>
<tr>
<td>617-629</td>
<td>Other disorders of female genital tract</td>
<td>0.0016</td>
<td>0.0641</td>
<td>102</td>
</tr>
<tr>
<td>640-649</td>
<td>Complications mainly related to pregnancy</td>
<td>0.0000</td>
<td>0.0000</td>
<td>86</td>
</tr>
<tr>
<td>660-669</td>
<td>Complications occurring mainly in the course of labor and delivery</td>
<td>0.0000</td>
<td>0.0000</td>
<td>51</td>
</tr>
<tr>
<td>680-686</td>
<td>Infections of skin and subcutaneous tissue</td>
<td>0.0285</td>
<td>0.0988</td>
<td>466</td>
</tr>
<tr>
<td>690-698</td>
<td>Other inflammatory conditions of skin and subcutaneous tissue</td>
<td>0.0003</td>
<td>0.0644</td>
<td>294</td>
</tr>
<tr>
<td>700-709</td>
<td>Other diseases of skin and subcutaneous tissue</td>
<td>0.1270</td>
<td>0.2938</td>
<td>667</td>
</tr>
<tr>
<td>710-719</td>
<td>Arthropathies and related disorders</td>
<td>0.0279</td>
<td>0.0655</td>
<td>677</td>
</tr>
<tr>
<td>720-724</td>
<td>Dorsopathies</td>
<td>0.0072</td>
<td>0.0387</td>
<td>485</td>
</tr>
<tr>
<td>725-729</td>
<td>Rheumatism, excluding the back</td>
<td>0.0164</td>
<td>0.0655</td>
<td>387</td>
</tr>
<tr>
<td>730-739</td>
<td>Osteopathies, chondropathies, and acquired musculoskeletal deformities</td>
<td>0.0676</td>
<td>0.1775</td>
<td>771</td>
</tr>
<tr>
<td>745-747</td>
<td>Circulatory system</td>
<td>0.0000</td>
<td>0.0007</td>
<td>352</td>
</tr>
<tr>
<td>780-789</td>
<td>Symptoms</td>
<td>0.0318</td>
<td>0.1036</td>
<td>3482</td>
</tr>
<tr>
<td>790-796</td>
<td>Nonspecific abnormal findings</td>
<td>0.0372</td>
<td>0.1182</td>
<td>959</td>
</tr>
<tr>
<td>797-799</td>
<td>Ill-defined and unknown causes of morbidity and mortality</td>
<td>0.0953</td>
<td>0.1847</td>
<td>218</td>
</tr>
<tr>
<td>800-804</td>
<td>Fracture of skull</td>
<td>0.0191</td>
<td>0.0375</td>
<td>423</td>
</tr>
<tr>
<td>805-809</td>
<td>Fracture of neck and trunk</td>
<td>0.0126</td>
<td>0.0415</td>
<td>774</td>
</tr>
<tr>
<td>810-819</td>
<td>Fracture of upper limb</td>
<td>0.0074</td>
<td>0.0297</td>
<td>324</td>
</tr>
<tr>
<td>820-829</td>
<td>Fracture of lower limb</td>
<td>0.0307</td>
<td>0.0703</td>
<td>353</td>
</tr>
<tr>
<td>830-839</td>
<td>Dislocation</td>
<td>0.1382</td>
<td>0.1382</td>
<td>94</td>
</tr>
<tr>
<td>850-854</td>
<td>Intracranial injury, excluding those with skull fracture</td>
<td>0.0383</td>
<td>0.0779</td>
<td>627</td>
</tr>
<tr>
<td>860-869</td>
<td>Internal injury of thorax, abdomen, and pelvis</td>
<td>0.0106</td>
<td>0.0278</td>
<td>600</td>
</tr>
<tr>
<td>870-879</td>
<td>Open wound of head, neck, and trunk</td>
<td>0.0067</td>
<td>0.0181</td>
<td>406</td>
</tr>
<tr>
<td>880-887</td>
<td>Open wound of upper limb</td>
<td>0.0000</td>
<td>0.0000</td>
<td>106</td>
</tr>
<tr>
<td>890-897</td>
<td>Open wound of lower limb</td>
<td>0.0000</td>
<td>0.0000</td>
<td>56</td>
</tr>
<tr>
<td>900-904</td>
<td>Injury to blood vessels</td>
<td>0.0000</td>
<td>0.0000</td>
<td>107</td>
</tr>
<tr>
<td>905-909</td>
<td>Late effects of injuries, poisonings, toxic effects, and other external causes</td>
<td>0.0000</td>
<td>0.0350</td>
<td>59</td>
</tr>
<tr>
<td>910-919</td>
<td>Superficial injury</td>
<td>0.0000</td>
<td>0.0255</td>
<td>58</td>
</tr>
<tr>
<td>920-924</td>
<td>Contusion with intact skin surface</td>
<td>0.0000</td>
<td>0.0159</td>
<td>136</td>
</tr>
<tr>
<td>930-939</td>
<td>Effects of foreign body entering through Body orifice</td>
<td>0.0854</td>
<td>0.1509</td>
<td>112</td>
</tr>
<tr>
<td>950-957</td>
<td>Injury to nerves and spinal cord</td>
<td>0.0300</td>
<td>0.0300</td>
<td>64</td>
</tr>
<tr>
<td>958-959</td>
<td>Certain traumatic complications and unspecified injuries</td>
<td>0.0429</td>
<td>0.0429</td>
<td>106</td>
</tr>
<tr>
<td>960-979</td>
<td>Poisoning by drugs, medicinal and biological substances</td>
<td>0.0068</td>
<td>0.0224</td>
<td>281</td>
</tr>
<tr>
<td>990-995</td>
<td>Other and unspecified effects of external causes</td>
<td>0.0479</td>
<td>0.1591</td>
<td>924</td>
</tr>
<tr>
<td>Code</td>
<td>Description</td>
<td>2011 Rate</td>
<td>2012 Rate</td>
<td>2011 Counts</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>996-999</td>
<td>Complications of surgical and medical care, not elsewhere classified</td>
<td>0.0364</td>
<td>0.1201</td>
<td>3880</td>
</tr>
<tr>
<td>E810-819</td>
<td>Motor vehicle traffic accidents</td>
<td>0.0034</td>
<td>0.0092</td>
<td>608</td>
</tr>
<tr>
<td>E820-825</td>
<td>Motor vehicle non-traffic accidents</td>
<td>0.0000</td>
<td>0.1293</td>
<td>56</td>
</tr>
<tr>
<td>E849</td>
<td>Place of Occurrence</td>
<td>0.0133</td>
<td>0.0370</td>
<td>488</td>
</tr>
<tr>
<td>E850-858</td>
<td>Accidental poisoning by drugs, medicinal substances, and biologicals</td>
<td>0.0302</td>
<td>0.0345</td>
<td>102</td>
</tr>
<tr>
<td>E870-876</td>
<td>Misadventures to patients during surgical and medical care</td>
<td>0.0583</td>
<td>0.0591</td>
<td>78</td>
</tr>
<tr>
<td>E878-879</td>
<td>Surgical and medical procedures as the cause of abnormal reaction of patient</td>
<td>0.0222</td>
<td>0.0315</td>
<td>1080</td>
</tr>
<tr>
<td>E880-888</td>
<td>Accidental falls</td>
<td>0.0367</td>
<td>0.0862</td>
<td>871</td>
</tr>
<tr>
<td>E910-915</td>
<td>Accidents caused by submersion, suffocation, and foreign bodies</td>
<td>0.0674</td>
<td>0.1033</td>
<td>66</td>
</tr>
<tr>
<td>E916-928</td>
<td>Other accidents</td>
<td>0.0180</td>
<td>0.0333</td>
<td>135</td>
</tr>
<tr>
<td>E930-949</td>
<td>Drugs, medicinal and biological substances causing adverse effects in therapeutic use</td>
<td>0.0296</td>
<td>0.0770</td>
<td>807</td>
</tr>
<tr>
<td>E950-959</td>
<td>Suicide and self-inflicted injury</td>
<td>0.0000</td>
<td>0.0132</td>
<td>194</td>
</tr>
<tr>
<td>E960-969</td>
<td>Homicide and injury purposely inflicted by other persons</td>
<td>0.0000</td>
<td>0.0000</td>
<td>131</td>
</tr>
<tr>
<td>V07-09</td>
<td>Persons with need for isolation, Other potential health hazards and Prophylactic measures</td>
<td>0.0741</td>
<td>0.1728</td>
<td>332</td>
</tr>
<tr>
<td>V10-19</td>
<td>Persons with potential health hazards related to personal and family history</td>
<td>0.0343</td>
<td>0.1025</td>
<td>2603</td>
</tr>
<tr>
<td>V20-29</td>
<td>Persons encountering health services in Circumstances related to Reproduction and development</td>
<td>0.0000</td>
<td>0.0000</td>
<td>53</td>
</tr>
<tr>
<td>V40-49</td>
<td>Persons with a condition influencing their health status</td>
<td>0.0358</td>
<td>0.1020</td>
<td>2281</td>
</tr>
<tr>
<td>V50-59</td>
<td>Persons encountering health services for specific procedures and aftercare</td>
<td>0.0363</td>
<td>0.0997</td>
<td>1162</td>
</tr>
<tr>
<td>V60-69</td>
<td>Persons encountering health services in other circumstances</td>
<td>0.1328</td>
<td>0.1373</td>
<td>222</td>
</tr>
<tr>
<td>V70-82</td>
<td>Persons without reported diagnosis encountered during examination and investigation of individuals and populations</td>
<td>0.0000</td>
<td>0.0000</td>
<td>79</td>
</tr>
</tbody>
</table>

5. Discussion

Our results confirm previous findings that LDA-derived topic features provide a promising boost to mortality prediction\(^1\). Although the features derived from the "vanilla" LDA achieve slightly higher AUC than Labeled-LDA, the “vanilla” LDA topics require domain experts to interpret the topics and associate them with the underlying disease representation. At the same time, our proposal of using Labeled-LDA model with ICD-9-CM codes as labels suggest a feasible way to achieve direct interpretability of topic features. Specifically, the top words of a topic derived with Labeled-LDA tend to be strongly associated with the corresponding definition of ICD-9-CM code. Note that expert evaluation of topic quality is also made easy by virtue of associating topics with ICD-9 definitions.

Our transparent predictive model effectively provides the ability to tailor mortality prediction to the particular diagnosis, with the Labeled-LDA topic model supplying an association score for each ICD-9-CM code via topic proportions. Several cancers are notorious causes of death as can be seen in Table 3. Likewise, the largest proportion of our patient cohort are cardiac patients, and Table 3 shows low average probability of mortality for the corresponding topics. On the other hand, dislocation was surprisingly ranked high in topic mortality. We examined the patients with dislocation and found that more than half of the patients were over 50 years old. It might suggest the poor recovery from dislocation of the elderly which causing following complications after they were discharged from the hospital.

We expect that using different methods of label selection to supplement frequency thresholding we used in this work may lead to improved prediction for the labeled LDA model. This may entail, for example, selecting the ICD-9 codes from specific levels of the ICD-9 hierarchy. This can be seen as similar to the topic granularity experiments in which the number of topics is changed in the regular LDA model.

Interestingly, the Labeled-LDA topic model can potentially be used to uncover relations between different diagnostic labels by virtue of examining the associated terms. As an example, some of top words in the topic associated with “Other metabolic and immunity disorders” such as ‘chest’, ‘artery’, ‘coronary’, ‘cabg’, and etc. may reflect the relationship between cardiac and metabolic diseases as described in Alvarez et al.\(^35\) and Naschitz et al.\(^36\). This suggests that using labeled LDA models which factor in the label frequency and interdependence, such as the ones proposed by Rubin et al.\(^37\) can potentially be used to explore the correlation between different labels.
6. Conclusions

We demonstrated the promising predictive power for 30-day and 6-month post-discharge mortality prediction using Labeled-LDA derived topic features. Because of the diversity and complexity of the diseases, our approach incorporated ICD-9-CM codes as knowledge input to guide topic model learning. Given an ICU record, the derived model could be used to determine the likelihood of post-discharge mortality and provide the physician with a justification for this assessment in the form of a combination of diagnostic codes associated with derived high-risk topics. In addition, ICD-9-CM topic features may be interpreted directly by healthcare professionals and patients for understanding the specific results of mortality prediction. In future work, different ICD-9-CM hierarchy and Labeled-LDA variants may be explored to improve the topic interpretability and prediction model.

7. Acknowledgements

This work was supported in part by a research grant from Philips HealthCare.

References

A First Step towards a Clinical Decision Support System for Post-traumatic Stress Disorders

Sisi Ma, PhD, Isaac R Galatzer-Levy, PhD, Xuya Wang, MS, David Fenyö, PhD, Arieh Y Shalev, PhD
New York University School of Medicine, NY, NY, USA

Abstract

PTSD is distressful and debilitating, following a non-remitting course in about 10% to 20% of trauma survivors. Numerous risk indicators of PTSD have been identified, but individual level prediction remains elusive. As an effort to bridge the gap between scientific discovery and practical application, we designed and implemented a clinical decision support pipeline to provide clinically relevant recommendation for trauma survivors. To meet the specific challenge of early prediction, this work uses data obtained within ten days of a traumatic event. The pipeline creates personalized predictive model for each individual, and computes quality metrics for each predictive model. Clinical recommendations are made based on both the prediction of the model and its quality, thus avoiding making potentially detrimental recommendations based on insufficient information or suboptimal model. The current pipeline outperforms the acute stress disorder, a commonly used clinical risk factor for PTSD development, both in terms of sensitivity and specificity.

Introduction

Chronic PTSD is distressful and debilitating\(^1\), following a non-remitting course in about 10% to 20% of trauma survivors\(^2,3\). The early response may provide sufficient information to identify individuals at high risk\(^4,5\) as multiple studies to date have identified numerous risk indicators of chronic PTSD, many of which are retrievable shortly after trauma exposure including early symptoms of PTSD, depression or dissociation, physiological arousal (e.g., heart rate), early neuroendocrine responses, gender, lower socio-economic status, the early use of opiate analgesics, the occurrence of traumatic brain injury, and a progressively growing number of genetic and transcriptional factors\(^6-10\). Despite these discoveries, the individual prediction of PTSD remains elusive, thereby leaving a major gap between scientific discovery and practical application.

One reason for such a gap is the current use of computational models that do not match the disorder’s inherent complexity in etiology. As attested by its numerous risk factors, the etiology of PTSD is multi-causal, multi-modal and complex. As such, the longitudinal course of PTSD reflects a converging interaction of numerous, multimodal risk factors. Moreover, specific risk markers and their relative weight can vary between individuals and traumatic circumstances. For example, head injury increases the likelihood of developing PTSD\(^11\) but does not occur to many survivors. Similarly, the contribution of female gender to the risk of developing PTSD varies between traumatic events\(^12\) and with specific genetic risk alleles\(^13\). One other challenge faced by practitioners is that the risk of developing PTSD often has to be determined based on incomplete data, due to the constraint of clinical resources. Thus, to accurately predict PTSD in individuals, one must account for complex and variable interactions between putative markers and develop a method that is robust to missing data.

Machine learning techniques are well suited for knowledge discovery and outcome prediction for diseases that have complex etiology and multifaceted manifestation like PTSD. Machine learning methods, especially supervised learning methods, can discover structures and interactions that are informative for prediction from high dimensional, multi-modal data\(^14-16\). However, the implementation of machine learning methods to clinical observations for individual level prognostics and diagnostics is limited in psychiatry.

In the current study, as an effort to bridge the gap between the existing academic knowledge about PTSD and individual level diagnostics and prognostics, we designed and implemented a clinical decision support pipeline to provide clinically relevant recommendation for trauma survivors. To meet the specific challenge of early prediction, this work uses data obtained within ten days of a traumatic event. The current pipeline outperforms the acute stress disorder, a commonly used clinical diagnostic criterion, both in terms of sensitivity and specificity. The decision support pipeline is interactive in nature. When the quality of the predictive model is unsatisfactory, the pipeline recommends more information to be collected for re-evaluation rather than making a potentially detrimental decision based on insufficient data.
Methods

Data

This study used data collected for the Jerusalem Trauma Outreach and Prevention Study (J-TOPS\textsuperscript{2,17}, ClinicalTrial.Gov identifier: NCT0014690). Participants were adults admitted to the emergency department (ED) following potentially traumatic events. Eligible participants were screened by short telephone interviews, and those with confirmed PTEs as per DSM-IV PTSD criteria A1 and A2 received structured, telephone-based interviews approximately ten days (9.6 ± 3.9 days) after trauma exposure. Participants with acute PTSD symptoms in the first assessment were additionally invited for clinical interviews, 29.5 ± 4.6 days after trauma exposure. Participants of the first clinical assessment were re-evaluated five, nine, and fifteen months after the traumatic event. For detailed procedures, see Shalev et al\textsuperscript{17}. Participants provided oral and written consent for all phases of the study. The study’s procedures were approved and monitored by the Hadassah University Hospital’s Institutional Review Board.

For the purpose of this study, we included individuals who had initial data available at ten days and at least two additional time points (n = 957). The initial traumatic event exposure included motor vehicle accidents (84.1%), terrorist attacks (9.4%), work accidents (4.4%) and other accidents (2.0%). Participants included in this study did not differ from those who were not included in gender distribution, age, general distress, initial PTSD symptoms and the frequency of exposure to new traumatic events during the study\textsuperscript{2}.

To characterize the stress response after the exposure to the traumatic events, PTSD symptom trajectory was constructed with latent growth mixture modeling (LGMM) on PTSD symptom severity measured at different times after the trauma (see figure 1 in reference 5). The LGMM identified three distinct stress response types/trajectories: non-remitting (17% of the participants), slow remitting (27% of the participants), and rapid remitting (56% of the participants). The goal of the study is to predict whether an individual will display a non-remitting stress response or a remitting stress response and recommend appropriate course of action using data collected in the ED and 10 days following the trauma.

To predict whether an individual will display a non-remitting stress response, 20 features measured in the ED and at 10 days after the trauma were selected as predictors. These variables were selected because they were shown to contain predictive information (measured by the area under the ROC curve) regarding PTSD symptom trajectory in a previous study\textsuperscript{5}. All the selected features can be measured at low cost during emergency room visits and through follow-up interview by phone, therefore, ideal for a clinical decision support system. The features used in the current study are listed in table 1.

Table 1. Features used to construct predictive models.

<table>
<thead>
<tr>
<th>Measurement time</th>
<th>Feature Name</th>
<th>Mean(SD)/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergency Department</td>
<td>Amount of Time in the ER (hrs)</td>
<td>3.51(5.19)</td>
</tr>
<tr>
<td></td>
<td>Age (years)</td>
<td>36.23(11.92)</td>
</tr>
<tr>
<td></td>
<td>Loss of Consciousness</td>
<td>9.62%</td>
</tr>
<tr>
<td></td>
<td>Tranquilizers</td>
<td>1.96%</td>
</tr>
<tr>
<td></td>
<td>Head Injury</td>
<td>29.61%</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle Accident</td>
<td>84.10%</td>
</tr>
<tr>
<td>10 days post trauma</td>
<td>Pain</td>
<td>6.49(2.59)</td>
</tr>
<tr>
<td></td>
<td>Patient: Clinical Global Impression</td>
<td>3.96(1.73)</td>
</tr>
<tr>
<td></td>
<td>Clinician: Clinical Global Impression</td>
<td>3.74(1.4)</td>
</tr>
<tr>
<td></td>
<td>Total Depression Score</td>
<td>16.61(5.63)</td>
</tr>
<tr>
<td></td>
<td>Total PTSD Score</td>
<td>9.41(3.63)</td>
</tr>
<tr>
<td></td>
<td>Nightmares</td>
<td>35.44%</td>
</tr>
<tr>
<td></td>
<td>Avoid thinking about reminders</td>
<td>72.62%</td>
</tr>
<tr>
<td></td>
<td>Social Support</td>
<td>71.97%</td>
</tr>
<tr>
<td></td>
<td>Derealization</td>
<td>33.75%</td>
</tr>
<tr>
<td></td>
<td>Scared</td>
<td>74.32%</td>
</tr>
<tr>
<td></td>
<td>Trouble Concentrating</td>
<td>59.05%</td>
</tr>
<tr>
<td></td>
<td>Reexperience of the event</td>
<td>39.59%</td>
</tr>
<tr>
<td></td>
<td>Want Clinical Support</td>
<td>50.51%</td>
</tr>
<tr>
<td></td>
<td>Worthlessness</td>
<td>2.03(1.59)</td>
</tr>
</tbody>
</table>

The Decision Support Pipeline
The goal of the proposed decision support pipeline is to assist practitioners in assigning proper treatment options for individuals who are recently exposed to a traumatic event. More specifically, upon receiving relevant information collected from an individual, the proposed decision support system would provide one of the following three recommendations: (r1) High clinical risk, recommend clinical monitoring or treatment. (r2) low clinical risk, no clinical monitoring or treatment recommended (r3) insufficient information to determine clinical risk, recommend more information to be collected.

The decision support pipeline consists of three components, illustrated in figure 1. The first component is a database containing data collected from individuals who were exposed to a traumatic event and followed for a period of 14 month such that their long-term stress response trajectory could be determined. In our implementation, we used data collected for the Jerusalem Trauma Outreach and Prevention Study (J-TOPS) as the database. Since the goal is the early detection of high risk individuals for PTSD, we included data collected in the emergency department and within 10 days after the trauma. The second component of the decision support pipeline is a prediction algorithm that models the relationship between the stress response trajectory and other information (demographic, clinical, psychiatric, etc.) in the database. The prediction algorithm produces a predictive model as well as quality metrics associated with the predictive model. In our implementation, upon receiving the information (features) about an individual, the database was queried such that the data for all individuals that have the same features measured were returned. The predictive model was built on this subset of the database. We used the support vector machine with radial basis function kernel as the predictive algorithm, and sensitivity and specificity computed from leave-one-out cross validation on the validation set as the quality metrics of the model. We consider predicting the individual would display remitting stress response when he/she would actually display non-remitting stress response ten times as costly as predicting the individual would display non-remitting stress response when he/she would actually display remitting stress response. This cost information was taken into account by the prediction algorithm when building the predictive model. The third component of the decision support pipeline is a recommendation generator. The recommendation generator generates recommendations based on the prediction and quality metric(s) computed by the predictive model. In our implementation, we used a simple decision rule: if the leave-one-out cross validation sensitivity (the first quality metric) is greater than or equal to and the leave-one-out cross validation specificity (the second quality metric) is greater than or equal to , make the recommendation according to the prediction made by the predictive model, otherwise make the recommendation that more information is needed to determine the stress response. . The quality metrics and the decision rule are critical in our pipeline. The quality metrics estimate the performance of the predictive model built for individual patients (consists of different features). And, the decision rule produces clinical recommendations only when the predictive model is determined to be high quality. This allows the algorithm to recommend potentially detrimental clinical decisions based on predictions that have low quality. The decision support pipeline was implemented in Python.

Figure 1: The workflow of the decision support pipeline. 1) A subset of data from the database is retrieved according to what features are available for the individual to be predicted. 2) A predictive model is built on the subset of the data by the predictive modeling algorithm. 3) Quality metrics is computed for the predictive model. 4) PTSD trajectory membership is predicted by the predictive model. 5) The decision generator generates recommendation based on the quality metric and the predicted trajectory.

Evaluation of the Decision Support Pipeline
Leave-one-out cross validation was implemented to evaluate the performance of the decision support pipeline, i.e. each individual’s PTSD trajectory membership was predicted using the model built with data collected from the rest of the individuals as the database. Three metrics were computed to evaluate the performance of the pipeline: (1) Percent of decisive decision: the percent of individuals that the pipeline is able to make a decisive decision on, i.e. recommending (r1) or (r2); (2) Sensitivity; and (3) Specificity. We evaluated the performance of the pipeline with 143 different sets of thresholds for the quality metrics (leave-one-out cross validation sensitivity and specificity computed with the data in the database).

We also calculated the predictive performance of acute stress disorder (ASD, defined according to the DSM-IV\textsuperscript{21}) measured around 10 days after trauma exposure. ASD is commonly used in the clinical setting to identify people who are at high risk for PTSD development in the initial period of time following trauma exposure.\textsuperscript{22}. ASD is a binary variable computed from ASD symptom measurements. The predictive performance of ASD was compared to the predictive performance of our pipeline. Sensitivity and specificity obtained from leave-one-out cross validation computed from the entire cohort was reported.

Results

Predictive Performance of ASD

In the current patient cohort, 17\% of the individuals developed non-remitting stress response. Predicting the post-traumatic stress response trajectory membership using 10-days’ ASD results in a sensitivity of 0.527 and specificity of 0.695.

Predictive Performance of the Decision Support Pipeline

The predictive performance of the decision support pipeline was evaluated with 143 sets of different threshold values (leave-one-out cross validation sensitivity and specificity computed on the validation set) for the quality metrics. The sensitivity, specificity and percent of decisive decisions of the pipeline corresponding to the 143 different threshold values were shown in figure 2. Not surprisingly, the percent of decisive decision decreases as the thresholds for the quality metric increases. On the other hand, sensitivity and specificity do not change linearly as the thresholds increase. In figure 2, the performance corresponds to sets of thresholds that result in more than 60\% of decisive decisions were highlighted with dashed line. Within that same threshold region, the range of sensitivity on the testing set is 0.616-0.667, and the range of specificity on the testing set is 0.697-0.726 (see figure 3 for sensitivity and specificity for individual threshold values, with the Pareto front highlighted). The performance of the decision support pipeline outperforms that of the ASD both in terms of sensitivity and specificity within this threshold region.

Figure 2: Model performance evaluated at different sets of quality metric (validation set sensitivity and specificity) thresholds. Three panels show percent decisive decisions, sensitivity and specificity on the testing set. Each square shows the performance of the pipeline under a particular threshold setting. The color of the square (different shades of red) represents the performance, with red mapped to the best performance. Black dashed line encloses threshold region where the model makes decisive decision for more than 60\% of the individuals. The most stringent threshold settings (top right corner of individual panel) did not result in any decisive decisions. These results were left blank in the figure.
Discussion

The main contribution of this study is the design and implementation of a decision support pipeline that automatically generates clinical recommendations for individuals who have recently been exposed to trauma. To the best of our knowledge, this pipeline is the first to provide individual level and personalized prognostics for trauma exposure with special consideration for the quality of the predictive model. The current pipeline outperforms the ASD, a commonly used clinical diagnostic criterion, both in terms of sensitivity and specificity. One distinctive characteristic of the decision support pipeline is the inclusion of quality metrics for the predictive models. The main purpose of the quality metric is to evaluate the quality of models constructed with different predictors, since in the clinical setting, not all predictor are available for all patients. If the quality of a predictive model (constructed with a subset of features) is inadequate, the pipeline recommends more information to be collected for re-evaluation, rather than making a decision based on the prediction generated by a suboptimal model. Some of the features used for prediction are routinely obtained in the emergency department, whereas other features can be obtained through phone interviews around 10 days after trauma exposure at low financial cost. Therefore, this decision support pipeline can be readily incorporated into a clinical setting with minimal modification of the existing clinical workflow. We are currently working on implementing software based on the proposed pipeline. The software would automatically integrate clinical observations from sources including the EHR, diagnostic questionnaires and clinical interviews, to generate appropriate recommendations. When the information regarding a patient is not sufficient for prognosis, the software would generate a list of additional information to be collected and re-evaluate when new information is available. We believe the current pipeline and the aforementioned software could be a valuable addition to a clinician’s toolkit, especially when insufficient information is available.

One direction to expand the current work is to test the performance of the pipeline in a different population. In the current study, the predictive model was constructed with the J-TOPS population and the evaluation of pipeline is conducted on the same population. This may result in optimistic performance estimation of the pipeline. Furthermore, a certain percentage of the J-TOP patients only have data collected in less than 2 time points and were exclude in the study, which might result in bias of the performance estimation. To address these issues, we are currently collecting data in the Bellevue hospital center in New York as an independent test set.

In the proposed decision support pipeline, all three components can be modified to further improve the predictive performance. The first component of the pipeline, the database for building the predictive model, can be constructed to suit specific diagnostic purposes. For example, if the goal is to predict stress response trajectory in the general population, including individuals from a diverse demographic and socio-economic background would likely increase the predictive performance of the algorithm. On the other hand, if the goal is to predict stress response trajectory in a specific population, for example, war veterans, it is more appropriate to restrict the database to the target population. Moreover, other variables that are potentially predictive of PTSD trajectory could be included in the database as potential predictive features of PTSD to improve the predictive performance and robustness of the prediction. Similarly, for the second part of the pipeline, any predictive algorithms can be implemented to generate the predictive model. If the database consists of predictive features for distinct domains (e.g. demographic, psychiatric measurements, clinical measurements and physiological measurements) utilizing ensemble methods might result in better predictive performance. In addition, the current implementation builds the predictive model based on subset of the database where the features present in the testing set are available. Exploring different data imputation methods may result in better utilization of information contained in the database, thus resulting in improved performance. Also, different methods for model performance estimation can be implemented. Instead of leave-one-out cross validation, cross validation or cross indexing can be utilized as performance estimation with limited computational resource. Lastly, for decision generator, the last component of the pipeline, one can expand the possible recommendations and design customized functions to convert the prediction of the predictive algorithms to actionable recommendations.

Conclusion

841
The current study demonstrated the feasibility of early individual level prediction for chronic PTSD. As a first step towards a decision support system for chronic PTSD, the proposed pipeline provides clinically relevant recommendations based on personalized predictive models. To avoid making potentially detrimental clinical recommendations based on insufficient information, the quality of the predictive model was computed and integrated into the decision support pipeline. The pipeline outperforms ASD, a criterion often used in the clinical setting.

Acknowledgement

Isaac Galatzer-Levy is supported by K01MH102415 from the National Institute of Mental Health. Arieh Shalev is supported by MH071651 from the National Institute of Mental Health. The authors are grateful to the high performance computing facility at New York University Langone Medical Center for providing computational resources for conducting the analytical experiments.

Reference

Reusable Filtering Functions for Application in ICU data: a case study

Vincent Major, ME1, Monique S. Tanna, MD1, Simon Jones, PhD MSc1, Yin Aphinyanaphongs, MD PhD1

1NYU Langone Medical Center, New York, NY

Abstract

Complex medical data sometimes requires significant data preprocessing to prepare for analysis. The complexity can lead non-domain experts to apply simple filters of available data or to not use the data at all. The preprocessing choices can also have serious effects on the results of the study if incorrect decision or missteps are made. In this work, we present open-source data filters for an analysis motivated by understanding mortality in the context of sepsis-associated cardiomyopathy in the ICU. We report specific ICU filters and validations through chart review and graphs. These published filters reduce the complexity of using data in analysis by (1) encapsulating the domain expertise and feature engineering applied to the filter, by (2) providing debugged and ready code for use, and by (3) providing sensible validations. We intend these filters to evolve through pull requests and forks and serve as common starting points for specific analyses.

1.0. Introduction

1.1. Motivation

Randomized-controlled trials are considered the gold standard in clinical research to determine the efficacy and adverse effects of an intervention. However, there are many limitations to performing a randomized-controlled trial, including financial and personnel resources as well as patient recruitment, especially if the condition or outcome being studied is not common. Similarly, prospective observational studies can be used to characterize a condition and its associated outcomes, but are time-consuming and require large populations when studying uncommon conditions. Well-conducted retrospective database analyses circumvent many of these issues given the large population of data available for research. Broader exploration of infrequent conditions and outcomes, or of specific cohorts is enabled by big-data analysis tools and retrospective analysis of databases such as MIMIC-III.

MIMIC is an openly available database developed by the MIT Lab for Computational Physiology1. MIMIC recently released an updated version, MIMIC-III, which contains de-identified patient records for >40,000 critical care patients between 2001 and 2012. MIMIC-III is a 26 table relational database connected by a combination of unique identifiers and dictionary tables that link each identifier’s definition and the corresponding data. For additional information on the data and tables contained within MIMIC-III please refer to the detailed documentation at http://mimic.physionet.org/gettingstarted/overview/.

Using MIMIC-III for analyses is challenging due to the complexity of the input data and the complexity of the imperative preprocessing steps. The complexity of the input data can require significant feature engineering and domain knowledge to get data into a state of analytic readiness. The complexity of the preprocessing steps can also require engineering efforts to reproduce and may lead to profuse bugs as others attempt to reproduce the work.

In this paper, we reduce this complexity by describing and open sourcing code functions which allow other researchers to apply sensible filters toward putting specific data types into analytic form. The functions reduce complexity. First, the functions aggregate domain expertise and feature engineering for each specific data type. Second, the functions are debugged and in active use in our analytic projects.

We expressly aim to publish functions according to several engineering standards. Our intent is to encourage modular sharing through our code and to promote reuse. Publishing functions, such as the ones we propose, have several advantages. First, they promote reuse. Other researchers may easily use our code. Second, the steps are justified. Researchers may examine our feature engineering choices in detail and make sensible choices for their own analyses. Third, they encapsulate domain knowledge by practitioners. This will allow researchers who do not have access to domain experts to apply sensible filtering to clean ICU data. Fourth, the functions are validated. We provide evidence that the filtering is reasonable. Fifth, they are debugged. The code is in active use for our ICU projects and has been tested. Sixth, they provide a baseline starting point for applying different kinds of preprocessing toward optimizing objectives. For example, researchers may use our filters as a baseline and then from these tweak and build their own filters.
1.2. Background – Sepsis-Associated Cardiomyopathy

We motivate our overall approach with a clinical example. This example drives the design choices that we have made along the way. Acute left ventricular dysfunction has been reported to occur in one third of critically ill patients\(^2\). The inflammatory state associated with severe sepsis and septic shock can lead to dysfunction of multiple organ systems, including myocardial depression resulting in sepsis-associated cardiomyopathy. Also known as septic cardiomyopathy, this condition is characterized by acute left ventricular dysfunction, reduced ejection fraction and normal or increased cardiac output. By definition, sepsis-associated cardiomyopathy is reversible with resolution of the acute septic state with early studies reporting resolution of left ventricular dysfunction in 7-10 days\(^3\).

Although cardiovascular dysfunction is known to increase mortality, the prognostic implications of sepsis-associated cardiomyopathy remains controversial. Whether these patients should be managed any differently from patients with sepsis without associated cardiomyopathy is also unknown. Data on the pathophysiology, management and outcomes of this entity are limited and the literature on this topic consists mostly of small observational cohort studies.

2.0. Methods and Results

2.1. General Approach

The general approach to this case study investigation was to sequentially filter a large, diverse ICU cohort to one that reflects patients indicative of a particular condition, in our case sepsis-associated cardiomyopathy. Simple, modular functions are written that can be modified and assembled in countless ways. We will present six functions in this case study that (1) describe identification of sepsis, (2) discern severity of cardiomyopathy, (3) identify sepsis manifestations on admission, (4) exclude CHF/cardiomyopathy in past medical history, (4) calculate volume of intravenous fluids administered, and (6) calculate a severity of illness score. The six functions fit broadly into filtering of the dataset or calculating clinically significant parameters. For each function, we describe our method and approach and provide a simple chart review or graphical validation.

MIMIC-III uses subject, hospital admission, and ICU stay identifiers to distinguish individuals and their distinct inpatient admissions to the hospital and ICU respectively. These identifiers serve as a key in the majority of tables within the database but cases of repeated admission within the database were observed, i.e. individuals were admitted to the hospital/ICU more than once. The decision to use HADM_ID as the primary identifier and translate into the others only when necessary was made since sepsis-associated cardiomyopathy is an acute but reversible condition and distinct hospital admissions are unlikely to be affected by previous admissions.

2.2. Filtering database to target cohort

2.2.1. Sepsis Identification

**Method:** The MIMIC-III database contains great depth and breadth of ICU data. In order to filter the database down to patients relevant for a sepsis-associated cardiomyopathy study, the first step was ICD-9 requirements for severe sepsis. Patients with severe sepsis, defined as evidence of the systemic inflammatory response syndrome, source of infection, and at least one organ function dysfunction, were selected using the discharge ICD-9 codes indicative of infection\(^4\) concurrent with ICD-9 codes indicative of new onset organ dysfunction\(^4,5\). The complete list of ICD-9 codes with the function are available at the project page: [https://github.com/vjm261/MIMIC-ICU-filter-modules/wiki](https://github.com/vjm261/MIMIC-ICU-filter-modules/wiki).

**Validation:** The ICD-9 codes were independently reviewed in 30 final cohort individuals. The raw discharge summary was reviewed and compared to the ICD-9 codes to determine if the ICD-9 codes truly reflect a severe sepsis cohort. In three of the 30 randomly selected patients, there was no identified infection during the hospitalization despite the presence of ICD-9 codes indicative of infection. This accuracy met the criteria for our study and we did not pursue further refinements of the coding.
2.2.2. Cardiomyopathy Identification (extracting ejection fractions)

**Method:** In addition to ICD-9 codes, for a patient to be included, at least one echocardiogram report must be contained within the database and contain a valid left ventricular ejection fraction (LVEF). LVEFs are searched for from the raw, unstructured, free-text echocardiogram reports in both numeric (e.g. LVEF > 55% or LVEF = 60-80%) and ordinal (e.g. normal or severe) formats. LVEF is characterized by four ordinal values (Table 1).

Table 1. Clinically defined values of left ventricular dysfunction and the associated numeric ranges of LVEF.

<table>
<thead>
<tr>
<th>LVEF numerical range (%)</th>
<th>Ordinal value</th>
<th>Text string examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVEF &lt; 30</td>
<td>Severe</td>
<td>severely depressed, severe LV/left ventric. systolic dysfunction, severe hypokinesis</td>
</tr>
<tr>
<td>30 ≤ LVEF &lt; 45</td>
<td>Moderate</td>
<td>moderately depressed, moderate LV/left ventric. systolic dysfunction, moderate hypokinesis</td>
</tr>
<tr>
<td>45 ≤ LVEF &lt; 55</td>
<td>Mild</td>
<td>mildly depressed, mild LV/left ventric. systolic dysfunction, mild hypokinesis</td>
</tr>
<tr>
<td>55 ≤ LVEF</td>
<td>Normal</td>
<td>low normal EF, LV function normal, normal LV/left ventric./biventricular systolic function, LV/left ventric. systolic function grossly preserved</td>
</tr>
</tbody>
</table>

Numeric values of LVEF are binned into the four ordinal values as there is no way of back-calculating the numeric LVEF from the ordinal value and the ordinal values are relevant to clinical decision-making.

Searching through the raw, unstructured, free-text notes to find the echocardiogram reports is one of the more computationally expensive tasks of the analysis. All of the free-text notes are included in one, 4 GB .csv file. A python-based command line tool called csvkit was used to extract only the echocardiogram reports and circumvent problems arising from a comma-delimited file containing commas within the text field. The echocardiogram reports were later searched in R using regular expressions.

**Validation:** Over 150 distinct echocardiogram reports were independently reviewed and compared to the LVEF value extracted. Several examples of various text phrases used to describe left ventricular dysfunction were identified and amended to the list (Table 2). Otherwise, we excluded all cases of failure to extract an LVEF. These reports either had no mention of the LVEF or were transesophageal echocardiograms (TEEs).

2.2.3. Admission diagnosis inclusion/exclusion

**Method:** Given that our ICD-9 inclusion criteria were derived from discharge ICD-9 codes, and there is no other way of determining the time of onset, we also incorporated admission diagnosis consistent with a manifestation of severe sepsis into our inclusion criteria. After this step, we assume that admission is a consistent zero-time point from which temporal data is linked and comparable among the cohort. The admission diagnosis is a free text-based field in the database that clinicians were required to populate. Unfortunately, as a result the quality of this field is often poor.

All admission diagnoses were split into basic components (separators of commas and semicolons) and each component with frequency ≥ 3 were independently reviewed to generate a list of inclusion and exclusion terms. Patients were included if their admission diagnosis field could be a manifestation of sepsis but excluded if it also contained any of an exclusion list (Table 2). The exclusions remove patients that are admitted primarily for a non-sepsis condition and includes accidents, fractures, heart failure, gastrointestinal bleed, stroke, trauma, and others. A group of patients failed to meet the requirements of either the inclusion or exclusion lists and were excluded for lack of information about their presenting diagnosis. Of the 7984 patients with severe sepsis and an interpretable echocardiogram, 4044 pass this admission diagnosis filtering step. The remainder were either actively excluded, n = 6, or do not have an admission diagnosis that resembles a manifestation of sepsis, n = 3934.
We chose to be strict in our inclusion criteria for admission diagnosis in order to achieve a more specific cohort at the expense of possibly losing some patients who may in fact have been admitted with severe sepsis. Furthermore, the use of text-based admission diagnosis could be greatly improved by restricting this field to pre-defined diagnoses rather than allowing a free text field, as there was a very skewed count distribution of admission diagnosis with hundreds of strings occurring only once or twice which were automatically excluded.

Table 2. Inclusion and exclusion lists for admission diagnosis.

<table>
<thead>
<tr>
<th>Inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>acute renal failure, adult respiratory distress syndrome, altered mental status, ards, arf, arrest, aspiration pneumonia, bacteremia, cardiac arrest, cardiogenic shock, cellulitis, change in mental status, cholangitis, cholecystitis, colitis, dehydration, diabetic ketoacidosis, diarrhea, diverticulitis, dyspnea, elevated lfts, elevated creatinine, empyema, encephalopathy, endocarditis, epidural abscess, failure to thrive, febrile neutropenia, fever, fib arrest, hepatic encephalopathy, hepatic failure, hepatitis, hyperglycemia, hypoglycemia, hypotension, hypotensive, hypothermia, hypoxia, infected graft, infection, infection, influenza, liver failure, meningitis, mental status change, mental status changes, methicillin resistant staph aureus, myasthenia, neutropenia, osteomyelitis, osteopenia, osteomyelitis, pancreatitis, pancytopenia, pea arrest, pna, pneumonia, post arrest, post op infection, pyelonephritis, tlo sepsis, renal failure, resp failure, respiratory arrest, respiratory distress, respiratory failure, sepsis, septic knee, septic shock, shock, shortness of breath, sob, spontaneous bacterial peritonitis, tachycardia, thrombocytopenia, unresponsive, urinary tract infection, urosepsis, uti, vomiting, weakness, wound infection</td>
</tr>
<tr>
<td>Exclusion</td>
</tr>
<tr>
<td>acute leukemia, acute subdural hematoma, blunt trauma, bowel obstruction, bright red blood per rectum, cerebrovascular accident, chf, chf exacerbation, congestive heart failure, fall, femur fracture, gastrointestinal bleed, gi bleed, head bleed, heart failure, hip fracture, intracranial hemorrhage, intraparenchymal hemorrhage, lower gi bleed, overdose, pulmonary embolus, pulmonary embolus, ruptured aaa, s/p fall, s/p motor vehicle accident, stemi, stroke, stroke-transient ischemic attack, stroke/tia, subarachnoid hematoma, subarachnoid hemorrhage, subdural hematoma, subdural hemorrhage, syncope, tia, transient ischemic attack, transient ischemic attack (tia), trauma, tylenol od, tylenol overdose, upper gastrointestinal bleed, upper gi bleed, variceal bleed</td>
</tr>
</tbody>
</table>

Validation: The resulting cohort’s raw admission diagnosis were independently reviewed in a random sample of 30 patients. Each individual’s discharge summary was read to determine whether the inclusion/exclusion criteria used adequately identified patients with severe sepsis. In three of the 30 randomly selected patients, patients had a condition other than sepsis on admission, but this was not discernable based on the admission diagnosis field. For example, one patient was admitted with a diagnosis of shortness of breath and initially thought to have possible pneumonia, but the clinical presentation was later thought to be consistent with volume overload without infection. All three of these patients were included in the cohort based on the ICD-9 inclusion criteria because they did in fact develop an infection later during the hospitalization. This accuracy met the criteria for our study and we did not pursue further refinements of the coding.

2.2.4. Past medical history

Method: The last cohort filtering step was inspecting the past medical history (PMH) within the discharge summaries for any mention of past heart failure, hypokinesis, ventricular dysfunction, or depressed LVEF. Patients with pre-existing heart failure and/or reduced LVEF cannot be reliably diagnosed with further sepsis-associated cardiomyopathy in addition to their pre-existing left ventricular dysfunction in this retrospective study. Identical to the echocardiogram reports, the discharge summaries are extracted from the raw table with csvkit and the PMH section of each semi-structured report searched with a list of heart failure related terms and then the aforementioned numerical LVEF algorithm (Table 3).
Table 3. Text strings and numerical LVEF used to exclude patients with past medical history of reduced LVEF.

<table>
<thead>
<tr>
<th>Text strings</th>
<th>Numerical LVEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>cardiomyopathy, depressed EF, depressed ejection fraction, depressed LVEF,</td>
<td>If a numerical LVEF existed, patients were excluded if LVEF &lt; 55%.</td>
</tr>
<tr>
<td>global hypokinesis, ventricular hypokinesis, ventricular dysfunction, heart</td>
<td></td>
</tr>
<tr>
<td>failure, HF, CHF, HFREF, HFPEF</td>
<td></td>
</tr>
</tbody>
</table>

Validation: The past medical history was independently reviewed in the discharge summaries of 30 final cohort individuals. The presence or absence of pre-existing heart failure, both systolic and diastolic, was accurately identified in all 30 randomly selected patients.

2.2.5. Resulting filtered cohort

At this stage, each patient in the cohort: 1) has severe sepsis, defined by ICD-9 codes for both major organ dysfunction and systemic infection, 2) had an echocardiogram during their admission with a recorded, and interpretable, LVEF measure, 3) is admitted to the hospital with an admission diagnosis consistent with a manifestation of sepsis, and 4) has no known history of heart failure or reduced LVEF. Patients that were not admitted to the ICU within 24 hours of their hospital admission (that is, they spent time on a ward before the ICU) were also excluded as their sepsis-associated cardiomyopathy was unlikely to be the cause of admission (n=795). Patients with zero records of fluid ins were also removed from the final cohort (n=48). This resulted in a cohort of n = 2350 patients without a prior history of heart failure who were admitted to a medical intensive care unit with severe sepsis and had an echocardiogram during their hospitalization. Figure 1 describes the number of patients in the cohort at each filtering step.

Figure 1. Summary of the filtering steps from the whole database to the final cohort.

2.3. Calculation of Clinical Parameters

2.3.1. Fluids

Methods: Given the lack of evidence on the best fluid management strategy in patients with sepsis-associated cardiomyopathy, one of our clinical questions was to describe the differences in volume of fluids administered among patients with varying degrees of left ventricular dysfunction and its association with mortality. To address the effects of fluid management on patient outcomes, the fluids given to the cohort must be calculated. MIMIC-III contains tables of the various enterally and parenterally administered products (“ins”) including intravenous saline, intravenous blood products, intravenous and oral nutrition and intravenous medications. Tables of “outs” include outputs such as urine and drain output. In this analysis, fluids are restricted to items with volume units (fortunately all recorded ins were consistently measured in mL, verified by manual inspection). For each patient, all ins were extracted and summed for the first 72 hours following admission to the ICU. Evidence behind early goal-directed therapy for sepsis supports
aggressive fluid resuscitation in the first six hours after admission\textsuperscript{7}, however, given that patients requiring continued ICU stay likely remained acutely ill, we analyzed fluids administered over the initial 72 hours to examine physician practices in this setting.

The fluid ins and outs are categorized by unique item identifiers (MIMIC-III: ITEM_ID), that arose from combining two systems, one of which allowed manual naming during data collection. An attempt was made to limit the fluid analysis to the fluid types most commonly administered for fluid resuscitation purposes, excluding nutrition for example, but the same type of fluid was listed in various different text strings, and the decision was made that fluids of any kind should have the same effect on fluid balance and therefore all types of intravenously or orally administered fluids were included.

Information from many different database tables is combined into one table that is then used for the analysis. The final table includes patient identifiers, length of stay in the ICU and hospital, gender, age, date of echocardiogram, LVEF, mortality flags, volume of fluids in and out over 72 hours, and the OASIS score.

**Validation 1 - Descriptive:** The validity of the fluids calculation is dependent on the accuracy and reliability of the individual recordings in the database. Some patients have zero or very few recorded fluid ins or outs over several days. Many patients do not appear to have been administered substantial fluids. These same patients are likely to be those who are less severely ill or ill in a manner where strict fluid management is less important, and it is likely that these patients’ ins were not being strictly recorded. The histogram of Figure 2 illustrates this highly skewed distribution with a large peak near zero representing the 55 (2.3\%) of patients with zero and 101 (4.3\%) of patients with less than 1000 mL of fluid ins recorded over 72 hours. Note that 48 patients with zero fluid records have already been removed for lack of information.

![Figure 2. Histogram of the fluids administered over 72 hours.](image)

**Validation 2 - External Cohorts:** In order to address the potential invalidity of the recorded fluid ins, comparison of the distribution of fluid ins and outs between patients with sepsis-associated cardiomyopathy and several other cohorts was made. Even if the absolute values of fluid do not reflect the actual quantity of fluids administered, if the relative fluid quantities are appropriate between cohorts, the analysis outcomes should remain valid.

The final filtering step of excluding patients with a past medical history of heart failure or reduced LVEF removes 736 patients that would otherwise be eligible for the main cohort. These patients will be the first group compared to the main cohort. Patients with previous left ventricular dysfunction will likely be administered less fluid to avoid fluid overload and pulmonary edema, despite their severe sepsis.

By considering the excluded group at each of the other filtering steps, three other cohorts are found. Severe sepsis patients, by ICD-9 diagnostic code, but without an echocardiogram or no recoverable LVEF (n = 6702), heart failure with no systemic infection, by ICD-9 codes (n = 6042), and systemic infection without major organ dysfunction, again
by ICD-9 code, \(n = 7079\). Finally, a completely unrelated cohort of active dialysis patients is included as a lower limit of fluid requirements \(n = 2417\).

Each patient’s fluid ins and urine output are summed over the first 72 hours of their ICU admission. An empirical cumulative distribution function is calculated for each group and overlaid in Figure 3.

![Figure 3](image)

**Figure 3.** Empirical cumulative distribution function of (a) fluids administered, and (b) urine output over 72 hours for each of six independent cohorts.

As expected, dialysis patients receive the least amount of fluid ins, and the least amount of urine output given that many are unable to produce urine. Similarly, heart failure patients without infection also receive a relatively small amount of fluid ins given their increased risk of volume overload and pulmonary edema. Patients with pre-existing heart failure and concomitant severe sepsis, however receive more fluid ins than their counterparts with heart failure without infection given that the management of sepsis includes fluid resuscitation despite the risk of volume overload. Patients with severe sepsis and no history of heart failure (main cohort) have the largest amount of fluid ins, again as expected given that fluids are the mainstay of sepsis management. These findings validate that, in aggregate, the fluids reflect our expectations in the various cohorts.

By a Mann-Whitney test, all fluids in medians are significantly different \((p < 0.05)\). For the fluids out medians, most comparisons are significantly different except comparisons between the heart failure with no infection, infection without organ dysfunction, and severe sepsis with no echocardiogram groups that have similar medians \((p > 0.05)\).

### 2.3.2. OASIS

**Methods:** Unfortunately, the absence of a reliable ICU severity of illness score in the database requires the analyst to calculate their own. APACHE II or IV require variables that are not easily obtainable or ubiquitous in MIMIC including numerous lab results. In addition, several parameters are difficult to address with limited retrospective data such as whether the patient has acute renal failure, other chronic organ insufficiency or is immunocompromised. Although these questions could be addressed with a combination of ICD-9 codes, DRG codes and text-mining through the discharge summaries, any attempt to address these factors would not be robust without significant optimization.

In order to control for differences in severity of illness amongst the cohort, we implemented the Oxford Acute Severity of Illness Score (OASIS)

As opposed to comorbidity scores, such as the Elixhauser score, which are often derived exclusively from ICD-9 diagnosis codes, the OASIS score is a severity of illness score optimized for predictive strength with the fewest, and easily obtainable, subset of APACHE IV data elements. OASIS includes the heart rate, mean arterial pressure (MAP), respiratory rate, temperature, pre-ICU length of stay, patient age, Glasgow Coma Score (GCS), urine output, if the patient is ventilated on day 1, and if the patient was admitted following an elective surgery. Similar to APACHE,

OASIS is implemented with the most extreme values of each variable and the total urine output over the first 24 hours in the ICU which reflects the time within which clinical decisions regarding fluid resuscitation are made.
Many of the required variables for OASIS were recorded and time-stamped in the database (MIMIC-III: CHARTEVENTS). Age is given (MIMIC-III: PATIENTS), pre-ICU length of stay was calculated as the difference between ICU and hospital admission (MIMIC-III: ICUSTAY and MIMIC-III: ADMISSION) tables, urine output was calculated similarly to fluid ins, ventilation was addressed by the existence of an FiO₂ value (MIMIC-III: CHARTEVENTS) and given our inclusion criteria there were no patients in our cohort who were admitted following elective surgery. This was validated by manually reviewing the discharge summaries of 50 randomly selected final cohort patients, none of whom were admitted after elective surgery.

The item identifier (MIMIC-III: ITEM_ID) for each OASIS variable was manually searched in the dictionary table (MIMIC-III: D_ITEMS), and then extracted from the very large, 29GB table (MIMIC-III: CHARTEVENTS). Searching through and extracting chart records (from MIMIC-III: CHARTEVENTS) in order to calculate the OASIS score was the most computationally expensive component of the analysis. Bash scripts were used to extract the manually derived item identifiers (MIMIC-III: ITEM_IDs) for each relevant parameter separately which were then further processed in R and later combined. Similar to the admission diagnosis or fluid in identifiers, the chart item identifiers are not systematically organized. Numerous codes for each parameter are included in the database, most of which are infrequently used but all are included. Please refer to the project page at https://github.com/vjm261/MIMIC-ICU-filter-modules/wiki for the complete list of ITEM_IDs used for each OASIS parameter.

The OASIS score is a sum of component weights for each parameter. That is, each parameter falls within a weighted bin and the combination of parameter scores is the resulting OASIS score sum. The maximum and minimum of each parameter were extracted for each patient and the greater OASIS score component was used to calculate the total OASIS score.

Several problems arose in the combination of OASIS components:

- Temperature existed in both Fahrenheit and Celsius; given that temperatures were usually recorded in Fahrenheit, that was searched for first and if no record existed, Celsius was converted to Fahrenheit.
- MAP was uncommon in CHARTEVENTS. Instead, MAP was calculated from blood pressure measurements that exist for every patient using MAP = (DBP*2/3) + (SBP*1/3), where DBP and SBP are the diastolic and systolic blood pressures respectively.
- GCS was recorded as a total score or in its components (eye, verbal and motor response); this was thought to be an artifact due to the change in the clinical information system during the data capture period, and the component scores were summed to result in a total GCS score. Although most patients had either the GCS total or all three GCS components recorded in the first 24 hours in the ICU, 17.2% of patients did not.
- No records of urine output existed for 11.0% of patients.
- No record of body temperature existed for 0.1% of patients.

The patients with missing GCS, urine or temperature were imputed with k-nearest neighbors (kNN) with k = 5. All of the OASIS parameters, including the min, max, and median of the repeated chart variables were used to impute total GCS, total urine and temperature min, max, and median. The percentage of patients with missing GCS scores who were also missing urine outputs was 12.1% (108 of 888 patients), which is similar to the percentage of patients missing urine outputs among the entire cohort (11%) and suggests that the missingness of data is uniform and not limited to a specific group of patients.

Validation: The OASIS score has been previously validated in a very large ICU dataset; however, we used logistic regression to validate our function used to calculate the OASIS score. The OASIS score was used with age and gender to predict ICU mortality. Age was factored into 10 year bins, referenced to the ≤ 30 years group, whereas OASIS was factored into five equal-population quintile bins with the least severe factor, OASIS ≤ 30, used as a reference point. Quintile bins were used as the proportion of patients that died and is consistent within each bin but different between bins. The odds ratio (OR) with a 95% confidence interval (CI) is presented with the corresponding p-value (Table 4).
Table 4. Logistic regression OR, 95% CI and associated p-value for each predictor variable.

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.08</td>
<td>[0.04, 0.15]</td>
<td>3.10 x 10^-13</td>
</tr>
<tr>
<td>OASIS (30,36]</td>
<td>1.79</td>
<td>[1.21, 2.67]</td>
<td>0.004</td>
</tr>
<tr>
<td>OASIS (36,43]</td>
<td>2.33</td>
<td>[1.62, 3.37]</td>
<td>5.60 x 10^-6</td>
</tr>
<tr>
<td>OASIS (43,49]</td>
<td>2.72</td>
<td>[1.88, 3.99]</td>
<td>1.80 x 10^-7</td>
</tr>
<tr>
<td>OASIS (49,70]</td>
<td>5.87</td>
<td>[4.14, 8.46]</td>
<td>2.60 x 10^-22</td>
</tr>
<tr>
<td>Age (30,40] years</td>
<td>1.09</td>
<td>[0.49, 2.51]</td>
<td>0.83</td>
</tr>
<tr>
<td>Age (40,50] years</td>
<td>1.15</td>
<td>[0.58, 2.43]</td>
<td>0.70</td>
</tr>
<tr>
<td>Age (50,60] years</td>
<td>1.21</td>
<td>[0.64, 2.47]</td>
<td>0.58</td>
</tr>
<tr>
<td>Age (60,70] years</td>
<td>1.39</td>
<td>[0.73, 2.82]</td>
<td>0.34</td>
</tr>
<tr>
<td>Age (70,80] years</td>
<td>1.48</td>
<td>[0.78, 3.01]</td>
<td>0.25</td>
</tr>
<tr>
<td>Age (80,90] years</td>
<td>1.48</td>
<td>[0.78, 3.04]</td>
<td>0.25</td>
</tr>
<tr>
<td>Age (90,100] years</td>
<td>1.13</td>
<td>[0.55, 2.47]</td>
<td>0.75</td>
</tr>
<tr>
<td>Gender (M)</td>
<td>1.14</td>
<td>[0.93, 1.40]</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Increasing OASIS score is significantly associated with an increased mortality rate among our cohort of severe sepsis patients, as expected. Relative to the lowest quintile of OASIS score (≤ 30), the odds ratios of the four higher quintiles are 1.79 [1.21, 2.67], 2.33 [1.62, 3.37], 2.72 [1.88, 3.99], and 5.87 [4.14, 8.46] respectively, supporting that OASIS calculated from the MIMIC-III database reflects the increased risk of death for more severely ill patients and validates the use of OASIS in both this database and others.

3.0. Discussion

3.1. Function Design

We applied several software design principles laid out in *Clean Code*\(^1\) such as input parameter reduction, descriptive parameter naming, and data abstraction to create simple and reusable functions. We also built functions at multiple levels of granularity. At the first level is to assume that the database input structure is fixed and then the filters are applied. At the second level is to have functions that only accept data primitives (i.e. strings). For example, two simplified functions in each category are:

```plaintext
extract_echoreports_from_notes_mimic_dbase(<MIMIC-III database>)
extract_ef_from_echos(<echo report>)
```

3.2. Generalization

In this report, we describe and publish several filters on ICU data. These functions are customized toward our specific example. However, we see our filters as routine data processing steps for producing features on a given output type. These filters may also serve as building blocks that may be customized to other examples. These other examples may cite and fork our work to build their own filter functions.

3.3. Data Structure Changes

Some of these functions are sensitive to changes in the underlying data structure. If the MIMIC tables change, then the code will need to be updated to take into account the modified structure.
3.4 Validations

For each filter, we provided a validation by either chart review or further analysis. For some cases, the accuracy metric was less than 100%. We could have iterated further in refining our criteria. However, our validations met our error tolerances and in the context of the sepsis-associated cardiomyopathy, we were able to continue our clinical analysis. Our validation methods are also biased toward charts that meet the filter requirements. For example, we do not consider patients that may have had sepsis but the ICD-9 simply wasn’t recorded. In analyzing these huge retrospective databases, we felt that missing these uncommon occurrences was acceptable for our clinical question.

Chart review validation focused exclusively on minimizing false positives (i.e. optimizing precision) to improve the quality of the filtered cohort. Further work is required to broaden each filter to improve recall and specificity. The optimal operating point, balancing precision and recall, is likely to be highly problem. For our use-case, maximizing the recall was not a priority, instead the filters presented ensure all patients meet all filtering conditions.

4.0 Conclusion

We open-sourced a series of filters and validations on complex medical ICU data. These functions encapsulate the domain expertise and feature engineering of the medical authors and serve as a starting point for other researchers to apply to their own modeling work. Through pull requests and fork processes through git, we expect these functions to evolve and serve as common starting points for analysis. We demonstrate filters for cohort selection, fluids, and calculation of an ICU severity of illness score. We also validate each filter through chart review and sensible graphs. The code repository is available at https://github.com/vjm261/MIMIC-ICU-filter-modules/wiki.

5.0 Acknowledgements

This work is supported by NIH grant 1UL1TR001445-01.

References

A platform for exploration into chaining of web services for clinical data transformation and reasoning

José Alberto Maldonado, PhD1,4, Mar Marcos, PhD2, Jesualdo Tomás Fernández-Breis, PhD3, Estíbaliz Parcero1, Diego Boscá, PhD4, María del Carmen Legaz-García, PhD3, Begoña Martínez-Salvador, PhD2, Montserrat Robles, PhD1
1Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Spain;
2Dept. of Computer Engineering and Science, Universitat Jaume I, Spain;
3Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, Spain;
4Veratech for Health SL, Valencia, Spain

Abstract

The heterogeneity of clinical data is a key problem in the sharing and reuse of Electronic Health Record (EHR) data. We approach this problem through the combined use of EHR standards and semantic web technologies, concretely by means of clinical data transformation applications that convert EHR data in proprietary format, first into clinical information models based on archetypes, and then into RDF/OWL extracts which can be used for automated reasoning. In this paper we describe a proof-of-concept platform to facilitate the (re)configuration of such clinical data transformation applications. The platform is built upon a number of web services dealing with transformations at different levels (such as normalization or abstraction), and relies on a collection of reusable mappings designed to solve specific transformation steps in a particular clinical domain. The platform has been used in the development of two different data transformation applications in the area of colorectal cancer.

Introduction

Electronic Health Record (EHR) systems are called to play an increasingly important role in health information technology. The quantifiable benefits mentioned in a 2006 AHRQ report include: savings from data capture and access; information connectivity for stakeholders; and decision support, improving efficiency, safety, and quality of healthcare. In many cases (e.g. information exchange, and interoperability with decision support systems), a key issue is dealing with the heterogeneity of clinical data. As a matter of fact, clinical data sources may differ in the data models, schemas, naming conventions and level of detail they use. Several works define some sort of virtual health record (VHR) to bridge these differences. In order to achieve standardization, a number of works base their VHR on a standard EHR architecture. Some authors take a step further and use clinical information models based on EHR standards for this purpose. These models could come in the form of openEHR archetypes, CEN/ISO EN13606 archetypes, Clinical Document Architecture (CDA) templates, or Detailed Clinical Models (DCMs). An important role of such models is providing structure and terminology-based semantics to data instances that conform to some EHR model.

Beyond this, the semantic web has been proposed as the natural technological space for the integration and exploitation of biomedical data. The semantic web describes a new form of web content meaningful to computers, in which the meaning is provided by ontologies. In the context of EHR systems, ontologies enable the formal representation of the entities and relationships involved in an EHR extract, and of the associated background knowledge. Linked Open Data (LOD) is the most prominent semantic web initiative to develop the Web of Data, in which datasets would be semantically connected over the Internet. LOD principles propose semantic publishing and sharing of data using RDF and OWL languages. Based on these contents, automated reasoners can be used to infer new information or to check logical data consistency.

In line with the above ideas, in a previous work we solved an EHR-driven phenotyping problem by means of a data transformation pipeline to convert EHR data stored in a proprietary database, first into information models based on openEHR archetypes, and then into RDF/OWL extracts which could ultimately be combined with an OWL domain ontology for automated reasoning. In the case study, an abstraction phase including certain calculations (e.g. counting, negation) was done at the archetype (or data) level, while the classification tasks were performed at the
ontological (or knowledge) one. The separation of concerns between the archetype and the ontology levels was identified as a key factor in achieving the interoperability goals for the project at issue. The decision will depend on the project itself as well as on the features of the clinical information models used and/or the ontology language chosen. On the other hand, we concluded that a wide range of interoperability scenarios could be envisaged. For example, in a particular clinical domain, different information models may be under consideration. As another example, the transformation to the ontological level may not be required at all, or it may be limited to a set of RDF triples representing the EHR extract.

In this context, this paper describes a proof-of-concept platform to explore alternative solutions to a range of clinical data transformation problems. The platform is built upon a number of web services dealing with the transformation at the different levels. For this purpose, it relies on a collection of reusable fine-grained mappings designed to solve specific transformation steps (e.g., openEHR archetype to RDF) in a particular clinical domain (e.g., colorectal cancer). The paper is structured as follows. First we elaborate on the services offered by the platform, which we have classified according to the kind of operations they perform on the data: normalization, abstraction, semantic publishing, and reasoning services. Next we describe the web-based platform that we have implemented based on these services to evaluate the feasibility and utility of the idea, and explain how it has been used in two different data transformation scenarios in the domain of colorectal cancer. Lastly, we present some concluding remarks and outline future work.

Web services for clinical data transformation

Normalization service

Dual model EHR architectures (EHRA) encompass a reference model and set of archetypes for modelling EHRs. The reference model represents the generic and stable properties of the EHR. Conversely, archetypes are used to define domain-specific information models such as blood pressure recording, discharge report or a lab result. Notice that any reference model, such as ISO 13606-1, openEHR reference model or HL7 CDA, may be used as basis for defining archetypes.

The normalization service deals with one of the main problems when adopting EHR standards: how to standardize existing data. This involves transforming EHR content into a data structure compliant with both reference models and archetypes. We face a problem known in the literature as the data exchange (translation or transformation) problem. This problem is a difficult one, since it deals with differences and mismatches between heterogeneous data formats and models. In the case of the EHR this problem is even more complex. On the one hand, we have the legacy data that conform to a particular schema, and with local semantics. On the other hand, we have EHR reference models and archetypes that have been defined without any regard to the internal architecture or database design of EHR systems. The main requirements on the target instances are that they shall be compliant with the target archetype, be non-redundant and be an accurate representation of the source data (EHR data to be normalized).

The output of this kind of services is an XML instance of the target schema (archetype). Our normalization process requires the source schema expressed as an XML Schema, the target archetype and a declarative mapping relating both schemas. The mapping language is based on tuple-generating dependencies (tgds). They are expressive enough to represent, in a declarative way, many of the schema mappings of interest. The tgds basically specify how to compute a value for an atomic attribute of the target schema (archetype) by using a set of atomic elements from the data source. In our setting these value correspondences are defined by a set of pairs, consisting of a transformation function and a filter. The simplest kind of transformation function is the identity function which copies an atomic source value into a target atomic attribute. This is the most common transformation function in normalization scenarios since often they only involve restructuring source data to make them compliant with the target schema. In those cases where it is necessary to transform source atomic values, a wide range of transformation functions are supported. As an example, Table 1 contains a simple mapping transforming gender codes. It transforms the local gender code in the source path /patient/gender into a normalized code (0 for male, 1 for female and 9 otherwise). Note that the order of the filter-transformation function pairs is relevant and only the first applicable function is executed, consequently the last filter acts as a ‘default’ condition.
Table 1. Example of mapping expression normalizing local gender codes.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Transformation Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>/patient/gender='M' OR /patient/gender='m'</td>
<td>0</td>
</tr>
<tr>
<td>/patient/gender='W' OR /patient/gender='w'</td>
<td>1</td>
</tr>
<tr>
<td>/patient/gender=0 OR /patient/gender=1</td>
<td>/patient/gender</td>
</tr>
<tr>
<td>true</td>
<td>9</td>
</tr>
</tbody>
</table>

Value correspondences allow us to hide much of the structural complexity of archetypes and reference model. Users do not need to thoroughly specify the logical relations between all the entities of the source and target schemas. It is only necessary to specify the navigation path of the attributes used in the mapping. The semantics of the data transformation is complemented by a default grouping semantics which can be customized in different ways to cope with complex mapping scenarios. Finally, taking into account the mapping specification, the archetype constraints, and the source schema, an XQuery script is generated. The script takes as input an instance of the source EHR data and generates an XML document that is compliant both with the archetype and the underlying reference model. As a main advantage, this approach makes possible to separate the mapping specification from its implementation.

For the specification of the mapping and the generation of the normalization XQuery script we rely on the LinkEHR platform\textsuperscript{12,13}. The normalization service has been implemented as a web service that takes as input the XQuery script and the source data. The actual data transformation is performed by an XQuery engine.

Abstraction service

While the normalization service focuses on the transformation of EHR data according to a standardized information structure, without significant modifications, the abstraction service deals with the transformations required when recorded data must be interpreted before use on the basis of clinical considerations. These transformations are generally based on specialized domain knowledge, and are often substantial in nature (e.g. the notion of severe comorbidity can be inferred from the existence of a total of 19 specific problems, but may not be recorded as such). This is what has been called inferential abstraction by some authors\textsuperscript{14}, and terminology abstractions or abstract definitions by others\textsuperscript{1}. Interoperability of EHR systems and decision-support systems is a typical scenario with such abstraction requirements.

This service, as the normalization one, deals with data transformations. The main difference lies in the fact that target instances do not need to be an accurate representation of source data. For instance, the target data may not be expressed at same granularity level as in the source data. As a consequence a wide range of data transformation functions are required such as arithmetical or string functions, type conversion, and especially aggregation or terminology abstraction functions.

This service has three inputs: the source schema expressed as an XML Schema or as an archetype, the target archetype (not necessarily based on the same reference model as the source schema) and the mapping relating both schemas. The mapping formalism is the same as in the normalization service; the only differences are the support for a broader set of functions in value correspondences (such as aggregation), and that archetypes can be the source schema of the mapping.

For instance, to obtain the value for the attribute “severe comorbidity”, assuming that the presence/absence of a number of morbidities affecting the patient can be obtained from the source schema, we could perform an abstraction based on the Charlson comorbidity score of those morbidities. Note that this is a rather complex example because the attribute “Charlson score” may not be included either in the source schema. The mapping expression for “severe comorbidity” will produce the value \textit{true} if the Charlson score is above a certain threshold, and \textit{false} otherwise. The filter part of the mapping will contain an adequate comparison expression whereas the transformation function will be a constant (either \textit{true} or false).

For the specification of conceptual abstractions over SNOMED CT we support the SNOMED CT Expression Constraint Language\textsuperscript{15}. This language, recently developed by IHTSDO, can be used for the definition of computable rules that define sets of clinical meanings represented by either pre-coordinated or post-coordinated expressions. For instance the constraint:
is satisfied only by disorders of lung which have an associated morphology of edema (or a subtype thereof). This constraint may be used to abstract from acute pulmonary edema or toxic pulmonary edema to pulmonary edema. Note that in contrast to mappings between raw EHR data and archetypes, mappings between archetypes have the potential of being reused as they are, since they often encode domain knowledge. We have extended the Archetype Definition Language (ADL) with support for this language for terminology binding.

The result of the mapping specification is again an XQuery script which takes as input an instance of the source schema and generates an XML document that is compliant with the target archetype. The LinkEHR platform is used both for the specification of the mapping and for the automatic generation of the XQuery transformation script. The abstraction service has been implemented as a web service and the data transformation is performed by an XQuery engine. We additionally use an execution engine for the SNOMED CT Expression Constraint Language (available at http://snquery.veratech.es) which is invoked by the XQuery scripts to perform the abstractions required.

Semantic publishing service

The objective of the semantic publishing service is to represent EHR data in RDF or OWL. The input to this service must be data coming from relational databases or XML EHR extracts. In the present work, we focus on the second case, so this service takes normalized XML EHR extracts as input, that is, the outputs of the normalization and/or abstraction services. By normalized we mean that the data must have been captured using archetypes and expressed in XML conforming to such archetypes. At this stage there are two possible output formats for the EHR data, RDF and OWL. The choice depends on the expected data exploitation use case. If the user desires to obtain a LOD-oriented representation, RDF will be the output format. RDF is oriented to semantic data representation, but with reasoning possibilities limited by the triple store used. OWL is the output format of choice for those scenarios in which automated reasoning is required. OWL is oriented to knowledge representation and its OWL DL level permits to use DL reasoning over the knowledge base.

The XML EHR extracts may not be sufficient to generate their semantic representation because they generally lack domain knowledge. Our transformation engine also uses (1) the OWL ontology that is used for creating the semantic content and acts as knowledge schema; and (2) the mappings between the archetypes used for representing the extracts and the OWL ontology, which define how the EHR data are transformed into OWL individuals. The mapping rules ensure that the EHR data are correctly transformed into the semantic format and prevent redundancy in the output dataset.

For instance, the mapping rules would describe how the information about an adenoma in an EHR extract should be represented in the OWL ontology. In this example, a rule could map the attribute size of the adenoma in the archetype to the corresponding datatype property in the OWL ontology that is associated with the class representing adenomas. Such rule would be systematically applied to all the EHR data instances to create OWL individuals of the adenoma class. This is an example of a basic mapping rule. Our approach includes three types of basic mapping rules, which permit to link archetype entities with OWL classes, datatype properties and object properties, respectively.

Our experience in semantic transformation reveals that the basic rules are not sufficient to get semantically rich datasets. Sometimes, we need to define rules that involve multiple input entities and one or more ontology classes. Besides, the data sources do not always include all the content that is needed to obtain a semantic representation of the data or, at least, such meaning is not made explicit either in the XML schema or in the corresponding table. Therefore, we need to provide additional information in the mapping rules to enrich the EHR data. This is solved in our approach by means of transformation patterns, which will generate additional OWL axioms. These patterns can be regarded as a set of OWL axioms that provides the semantics required and missing in the data source. The patterns are linked to archetype entities, allowing for a semantic transformation of data. Finally, the patterns are expressed in OPPL2 (http://oppl2.sourceforge.net). It should be noted that once the mapping rules are defined, they can be stored and reused in similar transformation processes. The function used in most of these mapping rules is identity, since the atomic values are not changed. The exception to this is the function to create URIs from atomic values. Besides, rules may have filters associated, which would only transform the data instances for which the conditions defined in the corresponding filter hold true.

The semantic transformation method is also able to prevent the duplicity of OWL individuals in case different XML extracts represent the same OWL individual. For this purpose, identity conditions are used, which define the properties that make an OWL individual unique. OWL semantics is applied in the transformation process, which
permits to ensure that only data consistent with the ontology constraints are transformed. Besides, when RDF is the output format, the process ensures that an RDF dataset consistent with the OWL ontology is produced. In our approach, the semantic publishing service is provided by the semantic web integration and transformation engine SWIT\textsuperscript{17} (http://sele.inf.um.es/swit).

Reasoning service

The previous services provide means for data transformation and abstraction, and their outputs are datasets that correspond to EHR extracts in formats like XML, RDF and OWL. The secondary use of EHR data is becoming increasingly important in clinical scenarios, because it enables the exploitation of EHR content with different healthcare-related purposes. One example of secondary use is the classification of patients (e.g. according to levels of risk). This involves applying a series of classification rules which partitions the patients in the groups of interest defined by the rules.

The input to this service is an EHR extract. The reasoning service also uses a set of classification rules, which are applied to the EHR extract. The reasoning service is able to produce two different outputs, which depend on the way it is invoked: (1) an EHR data extract in OWL format, which contains all the information inferred by the reasoner; and (2) the OWL classes to which the patient described in EHR data extract belongs. Hence, the first option returns the whole OWL content whereas the second one targets very concrete RDF triples. As an example of the latter, in the case of a colorectal cancer screening protocol, only the level of risk of a given patient would be returned.

Currently, our reasoning service has been implemented for EHR extracts in RDF/OWL format, that is, the result of the semantic publishing service. However, the solution is generic in the sense that the reasoning service could be implemented for EHR data coming in other formats such as XML. The current service deals with classification rules expressed in OWL. This decision is motivated by the possibility of applying DL reasoning over OWL content, so that state-of-the-art tooling can be reused for this purpose.

This service imposes two major requirements on the OWL ontology used for data classification. The first requirement is that it must contain one class per group of interest. Each group is implemented as an OWL class with equivalenceClass axioms, because they define the sufficient conditions for an OWL individual to be classified as a member of the class. This enables a DL reasoner to automatically partition the clinical data into the desired groups of interest. The second requirement is that the classification ontology must be aligned with the domain ontology used for representing the EHR data in OWL. The simplest way to achieve this is to define the groups of interest based on the domain ontology. Otherwise, the reasoner is still able to infer the corresponding classifications if equivalenceClass or SubClassOf axioms are established between the classes of the two ontologies.

Finally, this service has been implemented as two web services: one returns the OWL ontology with the inferred information, and another one returns the classification of a given individual. The reasoning of the input ontologies and the classification is obtained by applying the Hermit reasoner\textsuperscript{18}.

A proof-of-concept platform

In this work we have implemented a web-based platform that can be used to explore different alternatives for a clinical data transformation problem using the above web services. The platform works on a collection of reusable mappings designed for specific data transformation steps in a particular clinical domain. Based on the mappings, the platform offers a number of transformation alternatives or paths, made up of one or possibly more transformation steps, which can be then executed on sample data (see below for more details). The platform deals with this execution, by making the appropriate calls to the services involved (normalization, abstraction, etc.) and passing the output of each step to the input of the next one. Note that the possible transformation paths will vary depending on the available mappings. Also note that the mappings will determine the scenarios that can be explored (normalization only, normalization plus semantic publishing, etc.). The design of the platform is highly generic, so that new mappings can be easily added to explore different data transformation problems. At the time of this writing the platform includes a series of mappings related to the domain of colorectal cancer (see http://cliniklinks.upv.es/demoAMIA2016.html).

The following paragraphs detail the main steps when solving a data transformation problem with the platform, including the user interaction steps. They also describe the additional web services that have been implemented to support each step. The overall process and components are shown in Figure 1.
1) The platform displays automatically a list of all possible targets or destinations. These are retrieved from the currently available mappings through a specific-purpose web service. This action is performed in real time which facilitates the inclusion of new schemas and schema mappings. The user chooses the desired target.

2) Once the target has been established, another web service computes all possible transformation paths based on the mappings, in order to retrieve all the paths ending in the selected target. The mappings that define transformations between schemas and the schemas themselves may be distributed among servers in different organizations. The mappings and schemas can be interpreted as a directed acyclic graph, where the vertices are the schemas, and the edges are the mappings. A recursive in-depth search is performed which returns all the transformation paths from a source schema to the target one. The user selects which transformation paths to execute.

3) The user input for the execution of a transformation path is a clinical data instance based on the clinical data source schema. The instance may be provided directly from an EHR system or manually by clinicians by using digital data forms. Our proof-of-concept platform provides a web form based method to generate the input instance on the fly. The forms have been manually pre-designed according to the clinical data source schema. An additional web service automatically serves one form (each form is identified by the name of the first vertex in the selected transformation path) to the platform. The user fills out the displayed form to generate the source instance. In case multiple transformation paths are selected, the process would be much the same except for using multiple source instances (with different forms involved).

4) The platform includes a management web service that operates on the transformation paths and the data form contents. First, the management web service invokes another supporting web service that creates a valid source instance based on the form contents. With this instance, it then executes the corresponding web services for clinical data transformation (normalization, abstraction, semantic publishing, or reasoning transformations) in the required order. The management web service retrieves the output of each transformation step and hands it as input to the next one until the desired target is obtained, which is returned as result. It is also worth noticing that a single normalization or abstraction transformation may return a set of instances (e.g. in case the original source data contains information from more than one patient). This implies that all possible generated instances must be transformed so that further transformation steps are applied to each one of them.

**Figure 1.** Overview of the platform architecture along with the main steps to execute a data transformation scenario

**Application to the colorectal cancer domain**

We have used our platform for the development of two different data transformation applications in the area of colorectal cancer. Colorectal cancer is an important cause of death in developed countries. Screening for colorectal cancer serves to identify people who may be at risk of developing the disease. Concretely, the framework of our work is the colorectal cancer screening program of the Region of Murcia (Spain), which uses screening protocols standardized in Europe and America to classify patients according to their level of risk. As a result of the application
of these protocols, a database with data on more than 20,000 patients has been produced. Our goal is to demonstrate that the platform allows us to readily develop data transformation applications for different purposes, using EHR standards and semantic web technologies. In the first place, this requires a careful design of the different transformation steps, as well as of the models (archetypes, ontologies, etc.) to be used in each step, all this taking into account the needs of the project. Secondly, appropriate mappings must be defined for each transformation step, given the source and target of the transformation. Having defined the mappings, the platform will deal with the execution of the transformation pipeline, as described before. It should be noted that the mappings making up an application can be reused in other applications with similar transformation requirements.

In the rest of the section we describe the implementation of two applications using as starting point the database of Murcia’s colorectal cancer screening program. The first application corresponds to a basic normalization scenario, while the second is a more complex scenario implementing a colorectal cancer screening protocol. All the resources mentioned in the scenarios are available online (see http://cliniklinks.upv.es/demoAMIA2016.html).

Scenario 1: normalization to the CKM archetype for histopathology lab test results

This scenario corresponds to a single-step transformation performing a normalization. The input is the above database with the EHR extracts of the patients, and the output is a normalization of the previous database in terms of a standardized clinical information model. For this purpose we have chosen the openEHR-EHR-OBSERVATION.lab_test-histopathology archetype from the openEHR CKM repository.

Normalization

As stated before, the normalization service requires at design time the specification of a high-level mapping between the source schema and the target archetype. The source schema is a nested schema describing anatomic pathology studies with their findings. There is a top-level set of studies, and each study record has a nested set of test records. Tests have an additional nesting where each test has a set of findings. Studies and test records include context information such as dates and performers. Finally, findings contain the details of the microscopic findings such as size or pathological staging. In the target schema (archetype), there is a top-level record that represents a test. Each test has a set of microscopic findings and a set of macroscopic findings along with other information items such as specimen details or test status. Since the source schema has an additional nesting level (study), a single source instance might produce several target instances.

The mapping specification was created using the LinkEHR mapping capabilities. Since our purpose was to make public the available data in the form of an XML document compliant with the openEHR reference model, we did not employ any transformation function neither normalized the local terminology. It was not necessary either to define a mapping for all the atomic data elements in the archetype since a great part of them are not present in source data. Note that archetypes are defined to be as generic as possible in order to accommodate any relevant piece of information.

Scenario 2: implementation of a colorectal cancer screening protocol using abstraction, semantic publishing and reasoning

In this scenario we have used the platform to implement the case study of a previous work. The input to this clinical data transformation problem is the above-mentioned database with the EHR extracts of the patients, and the desired output is the level of risk for each patient. To take advantage of the platform’s capabilities for reasoning with OWL content, the EHR data will have to be transformed into OWL, to obtain the level of risk through classification reasoning. For this purpose, the semantic publishing service will be required.

Both the European and the American screening protocols define the level of risk by taking into account aspects like the amount of adenomas or the size of the largest adenoma, information which can be obtained from histopathologic findings. If this specific information is not included in the database, it will have to be calculated in terms of the services offered in the platform, that is, by means of either abstraction or reasoning services. For efficiency reasons, we consider more appropriate to perform operations like count or maximum as part of an abstraction step.

Hence, abstraction, semantic publishing and reasoning are the three services required in this scenario. Next, we describe the requirements and the inputs and outputs of each transformation step, including which inputs correspond to outputs of previous steps.

Abstraction
In the same vein as the normalization service, the abstraction one requires at design time the specification of a high-
level mapping between the source and target schemas, the only difference is that the source schema can be also an
archetype. In order to determine the patient’s risk level, we need to use concepts at two different levels of
granularity: at finding and at test levels. For the representation of finding-level concepts, we defined a specialization
of the archetype openEHR-EHR-OBSERVATION.lab_test-histopathology, named openEHR-EHR-
OBSERVATION.lab_test-histopathology colorectal_screening. This specialized archetype includes information
about adenoma findings such as the type or maximum size of the recorded dimensions (width, breadth and height),
the dysplasia grade and whether they are sessile and/or advanced. These information items are not stored directly in
the database but can be calculated by performing different types of abstractions. In order to represent test-level
concepts, concretely the maximum size of all adenomas and the number of adenomas, we developed a new
archetype from scratch called openEHR-EHR-EVALUATION.colorectal_screening.v1.

Two mapping specifications were necessary, the first one between the database and the finding-level archetype and
the second one between the finding-level archetype and the test-level archetype. The required abstractions were
codified in the mapping specification and therefore are performed by the resulting XQuery script. As explained
before, in our mapping formalism a value correspondence comprises a set of filter-transformation function pairs that
allows complex conditional expressions (if-then-elseif …). Value correspondences proved to be powerful enough
for defining the concept abstractions required in the first mapping (e.g. advanced adenoma). The second mapping
required additionally the use of aggregation functions: max for the calculation of the maximum size of all adenomas,
and count for the calculation of the number of adenomas. The archetypes and the mapping specifications were
created using the LinkEHR archetype editing and mapping capabilities, respectively.

Semantic publishing

The semantic publishing service uses the mapping designed to transform the EHR data conformant with the
archetypes of the previous step into the colorectal-domain OWL ontology, which provides the specific domain
knowledge. The ontology itself is also used in this transformation step. The service takes as input the EHR extract
produced by the abstraction service, and produces as output an OWL representation of that extract in terms of the
colorectal-domain ontology.

Reusing an OWL ontology from the NCBO BioPortal (http://bioportal.bioontology.org) was our first option to
promote the interoperability of the OWL content generated. We were not able to find any appropriate OWL
ontology, so we had to develop one specifically for this use case. The development of the colorectal-domain OWL
ontology was a manual process guided by the entities identified in the source database, in order to represent the
domain knowledge that would be required to express the meaning of the EHR extracts in OWL format. This
ontology includes classes for histopathology reports, types of findings, types of results of anatomical pathology
tests, etc., and properties and axioms that define precisely the meaning of these entities. The ontology consists of
102 classes, 47 object properties, 42 datatype properties and 1693 logical axioms.

The mapping file specifying the correspondences between the archetypes and the colorectal-domain ontology was
created using the SWIT tool. This mapping includes a transformation rule for each of the entities included in the
archetypes resulting from the abstraction step. Note that we have only defined the transformation rules for those
entities which we believe to be of interest for a semantic exploitation. Given that the mapping is defined between
archetypes and the domain ontology, it is independent of the source database.

Reasoning

The reasoning service takes as input the EHR extracts in OWL format resulting from the application of the semantic
publishing service. It operates by using the OWL classification ontology which contains the logical definition of the
levels of risk according to the European and American protocols. The output of the reasoning service is the level of
risk inferred based on both the input EHR extract and the classification ontology.

The OWL classification ontology was developed as an extension of the colorectal-domain one. Basically, the
classification ontology defines five classes, which correspond to the risk groups specified in the protocols. The
European protocol distinguishes high risk, intermediate risk, and low risk, whereas the American one only
differentiates between high risk and low risk. The five classes are modeled in a similar way. First, each class is
defined as a subclass of HistopathologyReport, which is a class of the domain ontology. Second, equivalentTo
axioms are associated with each class, which define sufficient conditions for membership of the class. For example,
the condition associated with both the LowRiskAmericanProtocol and LowRiskEuropeanProtocol classes is the
same, and is described as follows: “A histopathology report that only contains at most 2 normal adenomas, but does
not contain any advanced one". This is implemented in the OWL classification ontology as follows: HistopathologyReport and (hasAdenoma only NormalAdenoma) and (max_size some integer[< 10]) and (number some integer[< 3]). The classes of interest were mainly defined by using the properties and classes of the domain ontology. The classification ontology introduces 16 new classes and 62 new logical axioms, but no additional object or datatype properties.

Conclusion

In this paper we introduce the proof-of-concept platform that we have implemented to explore different alternatives for a clinical data transformation problem. It is built upon specialized web services dealing with data transformations which work at different levels, namely: normalization, abstraction, semantic publishing and reasoning. More importantly, the platform uses a series of reusable mappings that might be distributed across servers in different organizations and supports multiple EHR information standards (such as HL7 CDA, ISO 13606 or openEHR). In principle the mappings restrict the range of transformation applications (and hence scenarios) that can be configured using the platform. However, the design is rather generic, so that new mappings can be easily added allowing for a wider range of scenarios.

We also describe two different data transformation applications that have been implemented using our platform and web services. It is important to stress that the different models (archetypes, etc.) to be used along the transformation process, and specially the mappings defined taking into account the source and target of each transformation step, are essential in this kind of implementations.

Mappings involve local (XML) schemas, archetypes and ontologies as source or target elements. Concretely, mappings can be defined between: (a) an XML schema and an archetype, (b) an XML schema and an OWL ontology, (c) two archetypes, and (d) an archetype and an OWL ontology. As described earlier, we follow a “specify and generate” approach for all types of mappings. In this approach, developers are responsible for defining high-level mappings using specific-purpose tools, namely LinkEHR for mappings resulting in archetypes ((a) and (c)) and an ontology alignment format for mappings to RDF/OWL ((b) and (d)). These definitions are then automatically compiled into an executable script that will be used to perform the actual transformation. With mapping reuse in mind, it becomes necessary to pay attention to quality assurance aspects.

In our platform, automated reasoning supports both the semantic publishing and reasoning services, that is, the generation of the semantic dataset and its exploitation. State-of-the-art approaches do not use reasoning for guaranteeing the generation of consistent datasets, although reasoning has supported the transformation of clinical models into OWL. Those approaches mainly exploit reasoning as part of SPARQL queries or use specific rule languages such as SWRL, which are options more limited in terms of reasoning than OWL DL. Since we were interested in classification, OWL DL was deemed the preferred alternative. SPARQL and SWRL seem interesting options for other exploitation services that could be included in the future in our platform.

As future work we intend to extend our platform with different mapping management functionalities. Examples include the development of mapping operators, such as the merging of mappings, as well as of criteria (e.g. quality criteria in terms of the amount of information transferred) that can be used to compare mappings. Furthermore, we also foresee applications where the mappings (and/or models) listed in the platform are used in alternative ways. An example might be to use the mappings to determine the EHR data requirements for semantic publishing according to a given ontology. This will surely require including in the platform a precise characterization and/or a declarative description of the mappings. For instance, a formal description of the concept definition used in an abstraction mapping has proved very useful according to our previous experience. Finally, we plan to carry out the necessary developments with a view to the utilization of the data transformation applications built with our platform in a clinical setting.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness and the EU FEDER programme through grants TIN2014-53749-C2-1-R, TIN2014-53749-C2-2-R and PTQ-12-05620, and by Universitat Jaume I through project P1·1B2013-15.

References


Assessing Metadata Quality of a Federally Sponsored Health Data Repository

David T. Marc, PhD¹,², James Beattie, MLIS², Vitaly Herasevich, MD, PhD³, Laël Gatewood, PhD, FACMI², Rui Zhang, PhD²
¹The College of St. Scholastica, Duluth, MN; ²University of Minnesota, Minneapolis, MN; ³Mayo Clinic, Rochester, MN

Abstract

The U.S. Federal Government developed HealthData.gov to disseminate healthcare datasets to the public. Metadata is provided for each dataset and is the sole source of information to find and retrieve data. This study employed automated quality assessments of the HealthData.gov metadata published from 2012 to 2014 to measure completeness, accuracy, and consistency of applying standards. The results demonstrated that metadata published in earlier years had lower completeness, accuracy, and consistency. Also, metadata that underwent modifications following their original creation were of higher quality. HealthData.gov did not uniformly apply Dublin Core Metadata Initiative to the metadata, which is a widely accepted metadata standard. These findings suggested that the HealthData.gov metadata suffered from quality issues, particularly related to information that wasn’t frequently updated. The results supported the need for policies to standardize metadata and contributed to the development of automated measures of metadata quality.

Introduction

As part of the Open Government Initiative, the United States Federal Government published important datasets to the public using an online portal, HealthData.gov, to increase collaboration, transparency, consumer participation, and research. Despite the motivation to increase governmental transparency and accountability by publishing important healthcare datasets, there are considerable limitations to data access and retrieval. The lack of consistent standards to data storage and retrieval has been shown to inhibit citizen engagement in utilizing these resources for scholarly purposes. Martin, Foulonneau, and Turki conducted a study to examine if data that is considered open and public is actually accessible. They found limited openness of datasets advertised as open data, and considerable heterogeneity in the metadata elements that are intended to represent data sources. Shah and colleagues explained how publicly available biomedical data are annotated with unstructured text and are rarely described with ontology concepts available in the domain. The challenge is to adopt consistent metadata standards for public data to index the resources and allow for search methods that can identify the resource and related resources. Together, these issues challenge users in the discovery and access of public data. These limitations may be overcome with improvements to the quality of the metadata and adoption of standard annotation principles to index the publicly available data. Of particular interest is identifying the points of failure in the quality of the HealthData.gov metadata.

In HealthData.gov, metadata is used to summarize the available datasets and acts as the sole source of information for searching and retrieving relevant data. From 2012-2014, HealthData.gov metadata was maintained with the Comprehensive Knowledge Archive Network (CKAN), which is open-source data repository software. The CKAN platform requires federal agencies to submit specific metadata fields to describe a dataset (Table 1). Winn formally evaluated CKAN and found that there were limitations to the software including the inability to browse metadata history and workflow challenges around data storage and management. The advantages of CKAN include the flexibility for organizations to adapt the use of the platform with the opportunity to support data exchange using the Resource Description Framework (RDF) to facilitate data merging under a common schema.

Until now, the quality of storing publicly available healthcare data on HealthData.gov has not been formally evaluated. There are documented challenges around public participation of HealthData.gov, but the underlying causes of the challenges are not well explained. Therefore, the primary aim of this research is to formally examine the quality of the HealthData.gov metadata by assessing completeness, accuracy, and consistency to create standards and policies for publishing publicly available healthcare data.
### Table 1. Required metadata fields for HealthData.gov.

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>GUID</td>
<td>The unique identifier for the dataset</td>
</tr>
<tr>
<td>title</td>
<td>plain text</td>
<td>The title name for the dataset</td>
</tr>
<tr>
<td>notes</td>
<td>plain text</td>
<td>The description of the dataset</td>
</tr>
<tr>
<td>notes_rendered</td>
<td>HTML text</td>
<td>The description of the dataset rendered in HTML using Markdown</td>
</tr>
<tr>
<td>author</td>
<td>plain text</td>
<td>The name of the federal agency that submitted the dataset</td>
</tr>
<tr>
<td>url</td>
<td>url</td>
<td>The URL to the home page for the dataset, which links to downloadable files</td>
</tr>
<tr>
<td>tags</td>
<td>array of strings</td>
<td>Tags associated with the dataset</td>
</tr>
</tbody>
</table>

### Methods

The quality of the HealthData.gov metadata was evaluated using automated methods to assess completeness, accuracy, and consistency. The quality of the metadata was compared across the years the metadata was created and whether the metadata had undergone modifications at any point after the metadata was created.

#### Data Extraction

In August 2015, the complete HealthData.gov metadata catalogue from datasets that were published from January 2012 through December 2014 was queried from the CKAN engine using an HTTP GET Script. There were a total of 1,632 metadata instances extracted.

#### Completeness Assessment

The percentage of primary fields in each metadata instance that contained a null value was calculated to identify the frequency of blank fields. Additionally, an aggregated measure of completeness was obtained for each metadata instance. Equation 1 expressed how this metric was determined.

\[
Q_{\text{comp}} = \frac{\sum_{i=1}^{N} P(i)}{N}
\]

Where \( P(i) \) is 1 if the \( i \)th field has a non-null value, 0 otherwise. \( N \) is the number of fields defined in the metadata standard. The maximum value of this metric is 1 (in the case all the fields contain information) and the minimum value is 0 (all metadata primary fields are empty).

#### Accuracy Assessment

Accuracy is the degree to which metadata values are “correct” (i.e. how well they describe the object). Accuracy was assessed based on the proportion of working URLs and a semantic distance assessment.

The URL metadata field was classified as working or broken using a Python package that checks the syntax and connection of HTML documents. The proportion of error response codes was calculated. Only metadata instances where a URL was provided were included in the analysis.

Another measure of accuracy was the semantic distance between the information from the metadata instance and the information the same user could obtain from the resource itself. Formula 2 presents the semantic distance calculation, which was adapted from Ochoa and Duval\(^1\). The text for the data source was derived from the information contained on the website at the URL specified in the metadata.

All of the text was transformed to lowercase, punctuation was removed, the resulting white space was stripped, and common English words (i.e., stopwords) were removed. Two vectors were created based on the distinct words that appeared in the text of the original resource and the text of the metadata instance. The vectors were preprocessed before measuring the distance by removing any words that were not found in biomedical text using the biomedical dictionary GSpell\(^2\). GSpell is a spelling suggestion tool that uses a variety of algorithms to retrieve close neighbors. Any word identified as a misspelling was removed from the vectors. The reason this was done is to remove aberrant
words that may have resulted from text that was embedded within URLs found on the source page (e.g., .html, https).

The semantic distance was determined based on the comparison of terms that appeared in the metadata instance that also occurred in the original resource text. In Equation 2 the distance formula is presented.

\[
Q_{\text{accu}} = \frac{\sum_i \text{source}_i}{\sum_i \text{metadata}_i}
\]  

(2)

Where \(\text{source}_i\) is the binary output that indicated whether a word appearing in the metadata text also appeared in the source text of the \(i\)th dataset and \(\text{metadata}_i\) is the relative frequency of terms of the \(i\)th dataset. \(N\) is the total number of distinct words in the metadata text. The minimum value (lower quality) is 0, meaning that the text in the metadata does not exist in the source text. The maximum value (higher quality) is 1, meaning that all the words from the metadata appear in the source text.

To validate the \(Q_{\text{accu}}\) scores, 2 human reviewers independently assessed the level of agreement between the metadata and source text using a manual review process based on a method described by Ochoa and Duval\textsuperscript{11}. The metadata and the content from the URL were presented to the reviewers and the reviewers were asked to rate their level of similarity. The ratings were captured on a 7-point Likert scale from absolutely inappropriate to absolutely appropriate. The reviewer’s responses were assessed for reliability using an intraclass correlation coefficient (ICC). The human reviewer and \(Q_{\text{accu}}\) scores were transformed to categories where a score greater than or equal to 0.5 indicated high accuracy and anything less was low accuracy. Precision, recall, and \(F_1\) scores for \(Q_{\text{accu}}\) was assessed relative to the manual human reviewers as a method of measuring \(Q_{\text{accu}}\) accuracy.

**Consistency Assessment**

The consistency of the metadata was estimated as the degree to which the metadata matched standard definitions. There were two consistency measures: 1) Whether the fields of the metadata instances followed the CKAN requirements for the data type; 2) Whether the fields of the metadata framework followed the Dublin Core Metadata Initiative requirements.

The first consistency issue was measured by determining if each metadata field met the requirements of the data type as specified by CKAN\textsuperscript{9}. There were five different data types included in the metadata fields including GUID, plain text, HTML text, URL, and array of text. Rules were established to assess compliance with CKAN requirements, which are shown in Table 2. With the exception of the ‘Plain text characters’, the rules were assessed using regular expressions (regex) in the R statistical software package to determine if the format of the metadata field followed the CKAN requirements. The ‘Plain text characters’ rule was assessed by identifying the metadata instances that raised an exception to UTF-8 encoding. A logical vector was returned (i.e., TRUE or FALSE) indicating whether the regex found a match in the corresponding element in the input vector. The total number and proportion of TRUE values were added across each metadata field and for each metadata instance.

The consistency metric was based on CKAN compliance and equal to 1 minus the average of the fraction of problems found. Equation 3 was adapted from Ochoa and Duval\textsuperscript{11} and presents the calculation for the first consistency measure.

\[
Q_{\text{cons}} = 1 - \frac{\sum_i \text{brokenRule}_i}{N}
\]  

(3)

where \(\text{brokenRule}_i\) is a binary variable where 0 indicates compliance with a rule and 1 indicates noncompliance with a rule for each \(i\)th metadata instance. \(N\) is the number of rules in the metadata standard. The minimum value for the consistency metric is 0 (all possible errors were made) and the maximum value is 1 (there were no consistency problems).
Table 2. CKAN metadata requirements for data types that were evaluated.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Fields</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUID case</td>
<td>id</td>
<td>Are fields with GUID data type in lowercase?</td>
</tr>
<tr>
<td>GUID spacing</td>
<td>id</td>
<td>Do fields with GUID data type include hyphens in place of whitespace?</td>
</tr>
<tr>
<td>Plain text characters</td>
<td>title, notes, author</td>
<td>Do fields with a plain text data type only uses unicode characters encoded as UTF-8?</td>
</tr>
<tr>
<td>HTML characters</td>
<td>notes_rendered</td>
<td>Do fields with HTML text as the data type include two spaces for indentation and no trailing whitespace?</td>
</tr>
<tr>
<td>URL http</td>
<td>url</td>
<td>Do fields with the URL data type include a reference to a webpage indicated by the inclusion of “http”?</td>
</tr>
<tr>
<td>URL domain</td>
<td>url</td>
<td>Do fields with the URL data type specify the domain name?</td>
</tr>
<tr>
<td>URL page</td>
<td>url</td>
<td>Do fields with the URL data type specify the specific web page?</td>
</tr>
</tbody>
</table>

The second measure of consistency was determined based on compliance with the metadata requirements set forth by the Dublin Core Metadata Initiative\(^\text{13}\). The consistency measures evaluated the level of consistency between the HealthData.gov metadata field names and properties based on the Dublin Core standards. There were 9 rules evaluated and the frequency of compliance and violations was measured.

**Statistical Analysis**

The R Statistical Software package was used to carry out the statistical analyses. The proportion of URL errors that occurred in modified and unmodified metadata instances was compared using a two-sample Z-test of proportions. A three-sample Z-test of proportions was used to compare the proportion of URL errors across each of the three years the metadata instances were created. The Kruskal-Wallis statistical procedure with the Nemenyi post hoc test was used to test for differences in \(Q_{\text{comp}}\), \(Q_{\text{accu}}\), and \(Q_{\text{cons}}\) between each year the data was created. Between each year, the Mann-Whitney U test was used to compare differences in median values of \(Q_{\text{comp}}\), \(Q_{\text{accu}}\), and \(Q_{\text{cons}}\) based on the modification status.

**Results**

**Completeness**

When evaluating the frequency that required metadata fields included a blank entry, there were 0 for title, 19 (1\%) for author, 33 (2\%) for URL, 57 (4\%) for notes, and 157 (10\%) for tags.

As shown in Figure 1, the unmodified metadata had lower \(Q_{\text{comp}}\) scores compared to modified metadata (\(W = 154590.0\), p-value\(<0.001\)) and there were differences based on the year the metadata was created (Kruskal-Wallis \(\chi^2 = 13.5\); df= 2; p-value=0.002). Pairwise comparisons revealed that metadata created in 2012 had higher scores compared to 2013 (p=0.001) and 2014 (p=0.02). When comparing differences in \(Q_{\text{comp}}\) for each year the metadata was created and between each modification status, unmodified metadata had lower scores than modified metadata created in 2013 (W= 9924; p-value\(<0.001\)) and 2014 (W= 45750; p-value\(<0.001\)). Unmodified metadata created in 2012 had higher scores than unmodified metadata created in 2013 (W=641.5; p-value=0.006) and 2014 (W=2470.5; p-value=0.039). Modified metadata created in 2012 had higher scores than modified metadata created in 2013 (W=79062; p-value=0.005). Also, modified metadata created in 2013 had lower scores than modified metadata created in 2014 (W=144940; p-value=0.032)
Figure 1. Dotplot of the distribution and median values (red line) for $Q_{comp}$ scores by the year the metadata was created and modification status (n=1,632). Note: To prevent overplotting, the dotplot is presented with jitter. The scores do not exceed 1.

Accuracy

There were 1,601 metadata instances where a URL was provided and 71 (4%) resulted in an error. Of the 241 unmodified metadata instances, 32 (13%) had a URL error. There were 39 (3%) URL errors in the 1,360 modified metadata instances. There were significantly greater URL errors for the unmodified metadata compared to the modified metadata ($\chi^2=52.35$, df=1, p-value<0.001).

There were 29/290 (10%) errors for 2012, 6/579 (1%) errors in 2013, and 36/732 (5%) errors in 2014. The proportion of URL errors was significantly different across the three years ($\chi^2=37.38$, df=2, p-value<0.001).

The total number of 1,529 metadata instances were included in the semantic distance analysis. There were 103 instances removed due to invalid or missing URLs.

To validate the $Q_{accu}$ scores, 2 human reviewers independently rated metadata instances. The ICC statistic for reliability was 0.978, indicating strong agreement between raters. The precision of $Q_{accu}$ was 0.95 with a recall of 0.87 and an $F_1$ of 0.91. Together, these statistics verify strong accuracy of $Q_{accu}$.

As shown in Figure 2, unmodified metadata had significantly lower $Q_{accu}$ scores compared to modified metadata (W = 65899.0, p-value<0.001) and there were significant differences based on the year the metadata was created (Kruskal-Wallis $\chi^2$ = 368.04; df= 2; p-value=0.001). Metadata created in 2012 had a significantly lower $Q_{accu}$ scores compared to 2013 (p<0.001) and 2014 (p<0.001). The metadata created in 2013 had a lower scores than 2014 (p<0.001). When comparing differences in $Q_{accu}$ for each year between each modification status, unmodified metadata had lower scores than modified metadata for instances created in 2013 (W= 4457; p-value<0.001) and 2014 (W= 10308; p-value<0.001). Modified metadata that was originally created in 2012 had lower scores than modified metadata created in 2013 (W=24984; p-value<0.001) and 2014 (W=8062; p-value<0.001). Also, 2013 had lower scores than 2014 for modified metadata (W=75800; p-value<0.001).
Figure 2. Dotplot of the distribution and median (red line) of $Q_{accu}$ scores by the year the metadata was created and modification status (n=1,529).

Consistency

As shown in Figure 3, unmodified metadata had lower $Q_{cons}$ scores than modified metadata ($W = 162640.0$; p-value<0.001), but there weren’t differences based on the year the metadata was created (Kruskal-Wallis $\chi^2 = 6.11$; df= 2; p-value=0.05). Unmodified metadata had lower $Q_{cons}$ scores than modified metadata that originated in 2013 ($W= 8496$; p-value<0.001) and 2014 ($W= 50524$; p-value<0.001). The metadata created in 2012 had higher scores than 2013 for unmodified metadata ($W=633.5$; p-value=0.008) and lower scores for modified metadata ($W=721088$; p-value=0.031). Metadata created in 2012 had lower scores than modified metadata created in 2014 ($W=70961$; p-value<0.001). Unmodified metadata created in 2013 had higher scores than unmodified metadata created in 2014 ($W=3255.5$; p-value<0.001).

Figure 3. Dotplot of the distribution and median (red line) of $Q_{cons}$ scores by the year the metadata was created and modification status (n=1,632). Note: To prevent overplotting, the dotplot is presented with jitter. The scores do not exceed 1.
When comparing compliance with the Dublin Core standards, the metadata followed 3 of the 9 rules. Table 3 displays the rules that were confirmed and violated. The most common violations were related to the naming schema of fields.

**Table 3. Metadata compliance with Dublin Core standards.**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Definition</th>
<th>CKAN Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier field</td>
<td>Is the field representing the unambiguous reference to the resource labeled “identifier”?</td>
<td>No, the field is labeled “id”</td>
</tr>
<tr>
<td>Identifier content</td>
<td>Is the unique identifier identified by a string conforming to a formal identification system?</td>
<td>Yes, the string is made of 32 alphanumeric characters separated into the following segments 8-4-4-4-12</td>
</tr>
<tr>
<td>Title field</td>
<td>Is the field representing the name of a given resource labeled “title”?</td>
<td>Yes, the field is labeled “title”</td>
</tr>
<tr>
<td>Description field</td>
<td>Is the field representing the description of the resource labeled term “description”?</td>
<td>No, the field is labeled “notes”</td>
</tr>
<tr>
<td>Contributor field</td>
<td>Is the field representing the entity responsible for making contributions to the resource labeled “contributor”?</td>
<td>No, the field is labeled “author”</td>
</tr>
<tr>
<td>Date field</td>
<td>Do date fields use the encoding scheme W3CDTF profile of ISO 8601 (YYYY-MM-DDTh:mmTZD)?</td>
<td>Yes, the date field follows W3CDTF standards</td>
</tr>
<tr>
<td>Source field</td>
<td>Is the field representing the related source from which the desired resource was derived labeled “source”?</td>
<td>No, the field is labeled “URL”</td>
</tr>
<tr>
<td>Subject field</td>
<td>Is the field representing the topic of the resource identified by the label “subject”?</td>
<td>No, the field is labeled “tags”</td>
</tr>
<tr>
<td>Subject content</td>
<td>Does the field representing the topic of the resource represented by keywords from a controlled vocabulary?</td>
<td>No, the field is determined by the contributor and does not represent a controlled vocabulary</td>
</tr>
</tbody>
</table>

**Discussion**

The goal of this research was to investigate the quality of the HealthData.gov metadata. There was considerable growth in the number of datasets released each year, which suggested increased participation by agencies to publish important datasets in an open, accessible format. However, there was evidence of existing problems regarding the quality of the metadata representing the datasets, which is the foundational information of HealthData.gov. If problems exist in the metadata, this can impact the ability to search and retrieve data. To date, research has noted a lack of citizen participation in utilizing the public data, which has been attributed to inconsistent access standards. Until now research has not addressed the specific impact that inconsistent data standards have on the quality of the data.

Quality was measured based on the completeness, accuracy, and consistency of the metadata. These measures were chosen based on the recommendations of past research. Each measure of quality was compared between datasets based on the year the metadata was initially created and whether the metadata had undergone modifications at any point.

The $Qcomp$ was measured for all metadata instances and was calculated based on the proportion of fields that included data. There results of the analysis suggested growing issues around completeness. The $Qcomp$ scores were higher in 2012 compared to 2013 and 2014, suggesting more complete metadata in earlier years. Additionally, metadata that had undergone a modification at the time of the analysis had significantly higher scores (i.e., greater completeness) than the unmodified metadata.
Accuracy was measured by evaluating the frequency of broken links and the similarity of the metadata to the source data. About 5% of the URLs resulted in an error. Errors occurred more frequently in metadata that was left unmodified (13.28%) versus metadata that had undergone modifications (2.87%). Also, the metadata that was created in 2012 had the highest frequency of URL errors (10.0%) followed by 2014 (4.92%) and then 2013 (1.04%). The reason why there were more errors for 2014 when compared to 2013 is likely due to the greater number of unmodified metadata instances from 2014.

To better examine the accuracy of the metadata, an evaluation was conducted to determine how well the metadata represented the source data. The $Q_{accu}$ score is a quality metric that was calculated for each metadata instance and demonstrated how well the metadata represented the source data. When comparing the years the metadata was created and the modification status, it was found that $Q_{accu}$ scores significantly increased each year and the unmodified metadata had lower scores than the modified metadata. Together, these findings suggested that the modified metadata offered a better representation of the source data compared to unmodified metadata. Therefore, frequent updates to the metadata are recommended to ensure the metadata accurately represent source information.

Research supports this recommendation by showing improvements in the quality of the metadata with modifications$^{14,15}$. A final quality evaluation included a comparison of consistency. Consistency was measured based on the requirements specified by CKAN, the data portal where the HealthData.gov was published, as well as requirements defined by the Dublin Core, a standard for metadata design and infrastructure. The consistency metric, $Q_{cons}$, was calculated for each metadata instance based on the compliance the 9 CKAN. The observed trends suggested declining consistency overtime with unmodified metadata having significantly lower $Q_{cons}$ scores compared to the modified metadata.

The second consistency measure examined compliance with the Dublin Core standards for publishing metadata. Largely, the field names had violated the naming schema outlined in the Dublin Core, where only one of the six key metadata fields (i.e., title) complied with the standard. Regarding the formatting of the data, the metadata did conform closer to the Dublin Core standards. That is, the unique identifier and date were formatted appropriately. However, there was not a field representing the topic of the metadata using a controlled vocabulary. Together, these findings are concerning given the fact that the Dublin Core was developed to offer generic metadata standards and is the most widely adopted vocabulary for a Resource Description Framework (RDF). RDF is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model enabling resource discovery and cataloguing to enable users to deal with information with efficiency and certainty$^{16}$. Of particular interest is the role of a controlled vocabulary for representing the topic of metadata to help facilitate the search and retrieval of results. Although HealthData.gov included tags (i.e., keywords) as a required metadata field, these tags did not conform to a controlled vocabulary.

Given the evidence of quality concerns regarding the HealthData.gov metadata, there is greater need for open data policies to standardize a metadata framework and outline best practices for government participation. Following the publication of the Open Data Initiative, the U.S. Department of Health and Human Services (HHS) began to actively promote wide distribution of healthcare data through the Health Data Initiative (HDI). The Health Data Consortium (HDC) was developed to oversee the HDI and offer a structured approach to accelerate the availability and use of health data. The HDC is a public-private partnership that drives innovative uses of health information to improve health outcomes for individuals, enhance the effectiveness of the health care system, and build healthier communities. In October 2013, HDC created a strategic plan for the HDI, outlining the methods for advancing the distribution of health data$^{18}$. By making health data publicly available, the HDI intended to describe how the public can responsibly support healthcare innovations through creative, cost-effective, and efficient solutions$^{18}$.

Under HDC’s strategic plan, various metrics were developed to determine the level of achievement regarding established goals. The strategic plan suggested that if the results from these metrics were not meeting the set expectations, appropriate changes to the initiative would be put forth. Given the findings of this research, updated policies regarding standards for publishing metadata are required, which can include methods for assessing metadata quality.

In addition to updated policies, an updated metadata framework can help support greater quality, accuracy, and consistency. HealthData.gov uses a metadata schema called the “Project Open Data Metadata Schema”$^{19}$. This schema is extensible where metadata elements from Dublin Core can be included, however, Dublin Core is not the basis of the metadata schema. Rather, the Project Open Data Metadata Schema is based on the Data Catalog Vocabulary (DCAT), a hierarchical vocabulary specific to datasets$^{19}$. 871
An updated framework can be adopted to support higher quality metadata. To ensure complete metadata, a requirement can include mandatory fields. That is, a metadata instance cannot be submitted to HealthData.gov unless specific fields are completed. To address the accuracy concerns the metadata should be published using a standardized vocabulary for indexing purposes. In addition, the publisher of the dataset should verify the accuracy of the data at the time of the initial publication and on a quarterly basis. The system could incorporate a quality assurance process using an automated auditing tool to determine if submitted metadata meets a minimum quality requirement to ensure the information is complete, accurate, and consistent. Feasibly, the automated metrics used in this research could be adopted to assess metadata quality upon submission. To satisfy the consistency concerns, the metadata fields should follow the Dublin Core standards for the field names and formats. This includes not just the key metadata fields, but all metadata fields should abide by Dublin Core standards. An updated graphical user interface that offers contextual cues to support more complete and accurate data entry may help achieve greater consistency in adopting the aforementioned standards. ClinicalTrials.gov currently uses a quality audit where incomplete or inadequate submissions are denied and the submitter is informed of the identified issues. The submitter needs to address the issues with the submission before the data will be posted to the public.

At present, publicly available healthcare data is difficult to find and retrieve. The reason for such difficulties may be related to poor data quality and inconsistent use of standards. Policies and procedures regarding data storage that accompany the adoption of a data repository application may aid in addressing existing challenges. Future research should explore the application of a metadata schema that could facilitate greater quality of the metadata. Also, future research needs to be conducted to determine if higher quality metadata facilitates improvements in the retrieval and use of the publicly available healthcare data. Ultimately, this research can help to inform best practices for publishing metadata to inform policymakers and developers.

**Conclusion**

As part of the Open Government Initiative, the United States Federal Government published important datasets for use by the public. Health data is made available through an online data repository, HealthData.gov. Although there has been increasing participation by health agencies in disseminating data to the public, there has been very low public participation in the use of the data. A cause for the lack of public participation may be related to issues with the quality of the metadata that represents the health data. The metadata is the key source of information for storing and searching for datasets. Using automated methods of assessing data quality, clear issues regarding the completeness, accuracy, and consistency of the data were identified in HealthData.gov. There is a need develop policies that elucidate methods for publishing metadata to HealthData.gov that is complete, accurate, and consistent, which may include automated quality assessment, metadata standards, and terminology adoption.

**References**


A Data-Driven Approach For Better Assignment Of Clinical And Surgical Capacity In An Elective Surgical Practice

Gabriela Martinez, Ph.D. 1, Brian J. Bernard, M.S. 2, David W. Larson, M.D. 2, Kalyan S. Pasupathy, Ph.D. 1, Mustafa Y. Sir, Ph.D. 1, *

1 Mayo Clinic, Kern Center for the Science of Health Care Delivery, Health Care Systems Engineering, Rochester, Minnesota; 2 Mayo Clinic, Department of Surgery, Rochester, Minnesota

Abstract

This work analyzes strategies for better allocation of surgeon resources in an elective surgical practice. Among the metrics considered to evaluate the assignment of tasks are OR-to-Clinic ratio per provider, OR-to-Clinic ratio per day, patient access to clinic, and patient access to surgery. In addition, a simulation model is used to evaluate the clinical and surgical capacity of the calendar to identify potential inefficiencies and propose strategic changes to the calendar.

Introduction

The challenge of optimal assignment of human resources is one that is universal across all industries where human capital is a primary resource. In the health care industry this is especially true as people are involved in all aspects of patient care. From the person coordinating appointments, to the medical assistant checking vital signs, the doctor evaluating the patient, the nurse caring for hospital patients; the management decisions related to these resources are critical to a well-running integrated practice. Too few or too many of any of these resources can have negative impacts that could include: wasted resources, long waiting times, and poor quality of care. Health Care organizations are constantly challenged with these questions of resource assignment and how to best deliver quality care in the most efficient/cost effective way possible.

This paper addresses resource assignment in a specialty surgical practice with specific focus on how surgeons are assigned. In a specialty surgical practice, a surgeon can be assigned in one of two ways: clinic or surgery. Clinic time will result in new surgical procedures to be performed; while time assigned to surgery will reduce the number of procedures in the surgeon’s queue, see Figure 1.

Figure 1. General patient flow for an elective surgical practice.

It is important to maintain a proper balance of clinic and surgery. At a high level, too many patients entering the practice through clinic could result in very long queues for surgery. On the other hand, too much time assigned to surgery results in poor clinic access and under-utilization of operating room (OR) resources. These concepts are true of any surgical practice.

Capacity planning in health care industry is a complex problem which has been analyzed in the literature. For example, queue theory is used in 1 to determine the number of clinic appointments needed per day by a specialty clinic. The authors in 2 study the planning of operating rooms, they formulate an optimization problem to built a master surgical schedule. In contrast with previous works, this paper focuses on the strategic assignment of both clinic and OR. The surgeon assignment is done in a way to maintain appropriate OR-to-Clinic ratios for the overall practice as well as for each day in the planning horizon. The work presented in 1 considers a one-surgeon practice, our work considers time-allocation of multiple surgeons and its effects on patient access. As in 3, we use simulation to evaluate the performance of the calendar. The reader is referred to 4,5, 6 for a literature review on capacity planning problems in health care.
Materials and Method

This section discusses, in detail, an approach for addressing the surgeon time-allocation problem in an elective surgical practice at Mayo Clinic.

Assignment of Clinical Resources: Individual Calendar Pattern

Individual calendar patterns are the foundation for how surgeon resources are allocated between OR and Clinic. Mayo Clinic uses a 12-week planning horizon to assign clinical resources, surgeon calendars are updated on a rolling monthly basis. This means that calendars are available for clinical and/or surgical services twelve weeks into the future. Surgeons are assigned using a calendar where the days of the week are either blue or orange – in an alternating pattern. There are an equal number of blue and orange days each calendar year. A “blue” surgeon operates on blue days and is in clinic on orange days. This calendar structure creates a natural every-other-day clinic/surgery pattern for surgeons. Table 1 illustrates a two-week pattern for a surgeon. It can be observed that the resulting pattern for surgeon 1 defines a 1:1 OR-to-Clinic ratio.

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>OR</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Another scenario will illustrate a common calendar pattern for surgeons where non-practice time (A) is included, see Table 2. This time is generally related to research, education, or administration. The goal of the calendar patterns is to maintain a balanced practice for each surgeon where the number of new cases closely matches surgical capacity. For this to occur, the number of clinic days should be in the proper ratio to OR days. Therefore, non-practice time is added into the calendar pattern so that the resulting OR-to-Clinic ratio of the surgeon calendar is adequate. The placement of this time is based on the following criteria:

1) External requirements, such as, meetings/lab time on specific days
2) Surgeon clinic/OR flow
3) General practice clinic/OR flow.

Table 2 shows a calendar pattern of a surgeon with 20% non-practice time. The allocation of non-practice days into the calendar of surgeon 4 defines a 5:3 OR-to-Clinic ratio which is appropriate for surgeon 4 based on historical practice data.

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>OR</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

Assignment of Clinical Resources: Aggregate Calendar Pattern

The aggregate calendar pattern is the combination of all of the individual calendar patterns together. When we combine the eight individual calendar patterns, the resulting aggregate calendar pattern should be appropriate for the whole practice. This means that the OR-to-Clinic ratio is appropriate for each surgeon and also for each day of the two-week pattern. For this eight-surgeon practice the aggregate calendar with non-practice time is shown in Table 3.

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>OR</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>OR</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>OR</td>
<td>OR</td>
<td>C</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>OR</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>
Of the 14 non-practice (A) days assigned, 10 displace clinic days and 4 displace OR days. This results in an aggregate calendar that has: 1) an OR-to-Clinic ratio of 6:5 (36 OR and 30 Clinic days); 2) an even distribution of (A) days - one surgeon on (A) time on seven of the ten days; 3) a well-balanced daily OR-to-Clinic ratio.

Assignment of Clinical Resources: Aggregate Calendar Pattern with Discretionary Time

The calendar in Table 3 would repeat every two weeks if no other non-practice time were added. However, in reality there is additional discretionary time (D) that is available to all surgeons. This (D) time is made up primarily of trip time and paid time off. In this practice of eight surgeons there are approximately 500 days per year of (D) time – an average of about 10 days per week. This discretionary time, especially if the OR-to-Clinic ratio is changed significantly, can have a major impact on the surgeon and/or aggregate calendar. One of the policies in place to govern (D) time ensures a minimum of 4 surgeons assigned to either OR or Clinic each day. This is to ensure a minimum level of access for both Clinic and OR.

Table 4. Aggregate calendar with discretionary days added

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>M</th>
<th>T</th>
<th>W</th>
<th>Th</th>
<th>F</th>
<th>OR</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>D</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>D</td>
<td>D</td>
<td>OR</td>
<td>A</td>
<td>OR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>OR</td>
<td>OR</td>
<td>OR</td>
<td>C</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>A</td>
<td>OR</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>OR</td>
<td>A</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>OR</td>
<td>C</td>
<td>D</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>OR</td>
<td>C</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>OR</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>24</td>
<td>4</td>
</tr>
</tbody>
</table>

For the two-week period illustrated in Table 4, the addition of 19 discretionary days has a disruptive effect. As compared to Table 3: 1) Surgeon 1 has a ratio less than 1; 2) Surgeon 2 no longer has OR time; 3) the overall ratio has dropped below 1 (23:24); 4) three days (M, Th, and F) have an imbalance of OR and Clinic. These disruptions can have some of the following consequences:

- **Patient access** – too few clinic/OR days will result in poor clinic/OR access.
- **Day-to-day variation** – imbalance of surgeons assigned to clinic and OR (see Th and F) will cause variation in access and case volumes. Thursday will have high surgical volumes and low clinic access; Friday the exact opposite would occur.
- **Capacity utilization** – The overall ratio below 1 could result in any of the following: low utilization of clinics, reduced OR throughput, or OR overtime.

The management team recognizes that there is a need to monitor and adjust the practice calendar in order to minimize the impact of calendar disruptions. The next section discusses the process to strategically adjust the calendar to maintain a proper balance between OR and Clinic time.

Aggregate Calendar Planning and Maintenance Process

The management team meets each month to review and finalize the aggregate calendar. At the time of the meeting the calendar is finalized (patients can be scheduled) twelve weeks into the future. At the meeting, the calendar for weeks 13-16 is reviewed and finalized. Major changes can be made to the 13-16 week calendar and minor changes for weeks 1-12. An example of a major change is to switch a surgeon from Clinic to OR. Minor adjustments include additional resources assigned to a surgeon. This process is illustrated in Figure 2.
Figure 2. Calendar finalization process

At the management team meeting the aggregate calendar with discretionary days is reviewed along with the following practice data:

1) OR access - number of days to the next available day a case can be listed by surgeon and division
2) OR cases listed - total number of surgical cases listed in the 1-12 week period by surgeon and by division
3) OR-to-Clinic ratio per day
4) Number of surgeons assigned to clinic each day
5) Number of surgeons assigned to OR each day.

This data provides insight into the current state of the practice. Table 5 has an example of the data that would be used at this point in the process.

Table 5. Assignment allocation and practice data for surgeon 2 (13-16 week period).

<table>
<thead>
<tr>
<th>Clinic Assignment</th>
<th>OR Assignment</th>
<th>OR-to-Clinic Ratio</th>
<th>Cases Listed</th>
<th>Next Available OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 days</td>
<td>6 days</td>
<td>0.75</td>
<td>25</td>
<td>14 days</td>
</tr>
</tbody>
</table>

In Table 5, the cases listed and next available OR indicate that the OR schedule for surgeon 2 is currently busy with 25 cases listed and a 14 day wait for the next available OR. In addition, the OR-to-Clinic ratio is less than 1 for weeks 13-16 which is likely to further increase the queue for next available OR. This data suggests that major changes should be made to the schedule for surgeon 2; for example, a switch of one clinic to OR would increase OR capacity and bring the OR-to-Clinic ratio to 1. This analysis is done for each surgeon to identify opportunities.

The total number of surgeons assigned to clinic and OR on each day is also reviewed. As shown in Table 4, there can be imbalances in the number of surgeons assigned to clinic and OR on given days. In these instances, major changes are considered. The team looks for changes that can both positively impact the daily balance and improve the surgeon calendar(s). For example, in Table 4 for Surgeon 6 a major change from clinic to OR can be made on the second Friday. This change will balance the number of surgeons assigned to clinic and OR on that day. Also, it will change the surgeon 6 OR-to-Clinic ratio to be greater than 1. Any change to any calendar is allowed using this process as long as non-practice (A) time, approved discretionary (D) time, and more importantly, patient care are not impacted.

Performance of Calendars

This calendar management process has made a positive impact on the overall practice: OR capacity has been added with limited impact to clinic access (Figure 3). Anecdotally, we have seen reduced variation in the number of cases per day and per week, and the hospital census has become less variable. Table 6 shows the number of OR and Clinic days for the practice. The left columns reflect the calendar data before the above-described process was completed. For the same time period, the right columns reflect calendar data post-management process. Over this 9-month period, 24 additional days of OR capacity were added.
One of the counter-measures to the additional OR capacity is clinic access. The “third next available” metric is used to track the number of days from a given day to the third open clinic consult slot on the aggregate calendar. Figure 3 shows this metric for the same time period as Table 6. The average for this metric was 2 days which is considered acceptable clinic consult access.

In the process, there are a number of factors that are not considered that could yield more prospective and precise schedule changes. Patient arrival data, including variation of demand, is an important input to identify bottlenecks and periods of high variation. The performance of the schedule has many downstream impacts, for example, hospital census and staffing levels. The manual management process as described above can be automated to incorporate this additional data and metrics. The next section of this paper describes this more automated approach using simulation.

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>OR</th>
<th>Clinic</th>
<th>%OR</th>
<th>OR</th>
<th>C</th>
<th>%OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>54</td>
<td>53%</td>
<td>61</td>
<td>53</td>
<td>54%</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>41</td>
<td>47%</td>
<td>41</td>
<td>37</td>
<td>53%</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>38</td>
<td>52%</td>
<td>44</td>
<td>34</td>
<td>56%</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>58</td>
<td>50%</td>
<td>61</td>
<td>55</td>
<td>53%</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>44</td>
<td>58%</td>
<td>66</td>
<td>39</td>
<td>63%</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>83</td>
<td>49%</td>
<td>85</td>
<td>78</td>
<td>52%</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>20</td>
<td>60%</td>
<td>30</td>
<td>20</td>
<td>60%</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>74</td>
<td>50%</td>
<td>78</td>
<td>71</td>
<td>52%</td>
</tr>
<tr>
<td>Totals</td>
<td>442</td>
<td>412</td>
<td>52%</td>
<td>466</td>
<td>387</td>
<td>55%</td>
</tr>
</tbody>
</table>

Figure 3. Third next available clinic.

Data-Driven Analysis of Calendars

A model is used to evaluate various metrics into the decision making process of the calendar. The simple patient flow in Figure 1 is expanded to include metrics and downstream impact, for example, patient arrival and census. To this end, historical billing data of clinic appointments and surgeries performed by the practice were used to model patient flow, see Figure 4.
Four years of clinic appointment and surgical billing data were used to determine the surgical yield of the practice. The surgical and clinical data was linked by patient, then it was assumed that a clinic appointment within two weeks of the surgery date was the one which yielded surgery, more details can be found in 10. The estimated surgical yield of the practice is 68%. The proportion of inpatients, 70%, was estimated with billing data of surgeries performed in 2015. Patient arrivals were modeled with historical billing data on high-yield (HY) clinic appointments of 2015 since nearly 80% of patients of this practice are residents of other countries or other states of the US10. HY visits comprise clinic consultations of new patients (60 minutes visits) and comprehensive consultations (45 to 50 minutes visits) of established patients; these visits capture patients work-up/itineraries since early/mid week appointments are preferred to avoid weekend stays before surgery. Patients requiring short clinic visits (15 minutes) are not considered in the model In what follows, we provide more details of the parameters used in the simulation.

- **Patient demand (random)** – Daily patient arrivals were modeled as homogeneous Poisson processes11 with intensity parameters estimated from (HY) visits performed by the practice per day. The estimated patient arrival rates for the practice are shown in Table 7. The arrival rate per day shows that high-request days are Tuesday, Wednesday, and Thursday. The arrival rate of Fridays are low to accommodate travel patterns of national/international patients.

Table 7. Poisson process daily arrival rates.

<table>
<thead>
<tr>
<th>Arrival rate M T W Th F</th>
<th>( \lambda_d ) [patients/day]</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>10.04</td>
</tr>
<tr>
<td>T</td>
<td>16.96</td>
</tr>
<tr>
<td>W</td>
<td>12.07</td>
</tr>
<tr>
<td>Th</td>
<td>15.68</td>
</tr>
<tr>
<td>F</td>
<td>7.92</td>
</tr>
</tbody>
</table>

- **Clinical capacity (deterministic)** – Clinical capacity of a surgeon is defined as the historical maximum number of HY visits seen per day; the parameters are listed in Table 8. The clinical capacity of the calendar depends on the combination of surgeons assigned to Clinic. For example, the calendar in Table 4 has a clinic capacity of 19 and 13 consultations on Thursdays which are high patient arrival days.

Table 8. Daily clinical capacity per surgeon. Number of appointment slots available per day.

<table>
<thead>
<tr>
<th>Day</th>
<th>Surgeon</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td></td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>7</td>
<td>13</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>T</td>
<td></td>
<td>10</td>
<td>15</td>
<td>9</td>
<td>14</td>
<td>7</td>
<td>11</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>10</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>13</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Th</td>
<td></td>
<td>11</td>
<td>13</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>11</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

- **OR capacity (random)** – Historical data on surgeries performed in 2015 was used to estimate the number of cases that a surgeon can perform on an OR day. As illustrated in Figure 5, there exists a difference among surgeons, which implies that the surgical capacity on a given day depends on the combination of surgeons assigned to the OR.
• **Census (random)** - The length of hospital stay (LOS) of an individual post-surgery was estimated with admission/discharge data of 2014 and 2015. The staffing level for postoperative care for this practice is generally constant, therefore, staffing levels effects were ignored. The average length of stay is 5.04 days; the discrete LOS probability distribution is shown in Figure 6.

The simulation models the interaction between clinic and OR calendar days considering patient arrivals. For each calendar day the number of patients requesting a clinic visit is generated with its corresponding weekday-arrival rate parameter, and OR capacity is sampled from the distributions of the surgeons assigned in OR. It is assumed that patients prefer to have their clinic appointment close to their arrival date, then the following metrics are estimated: 1) access to clinic - the number of days patients need to wait for a clinic consultation; 2) patient seen – number of patient seen in clinic; 3) new surgical cases – surgical yield of each patient seen in clinic is 68%; 4) under/over-utilization – unfilled or fully utilized daily clinic capacity. The OR capacity and the number of patient seen in clinic are used to estimate: 5) access to OR - number of days a new surgical case needs to wait to access OR; 6) number of surgeries performed; 7) census – admission rate is 70% and length of stay is sampled from the LOS distribution.
Results

The simulation model was used to evaluate the performance of the aggregate calendar for the period of October 2015, November 2015, December 2015, and January 2016. The simulation was conducted after the placement of discretionary (D) days and changes to OR-to-Clinic ratio were made. Figure 9 shows the number of patients seen in clinic and their access to clinic. The left figure in Figure 9 shows the simulated patient seen in clinic, the dotted red line is the clinical capacity of the calendar and the percentile lines corresponds to the number of patient seen in clinic with probability 0.50, 0.75, and 0.90. It can be observed that the current clinical capacity is conservative resulting in unused appointments on several days. Most of the days with this clinic time waste have 3 to 4 surgeons assigned in clinic. The figure on the right shows the average time to access clinic (blue line) and worst-case time observed in the simulation (red line); the results show that third available clinic access of the calendar is at most three days.

Figure 9. Left figure: Simulated patient seen in clinic. Right figure: access to clinic.

The simulation results suggest that a clinical capacity level equal to the 75-percentile of patients seen in clinic could be an adequate level for the practice without compromising clinical access. To validate this, we compared the suggested clinical capacity level (75-percentile of Figure 9) with the historical number of clinic consultations of the period analyzed. Figure 10 shows that the suggested clinic capacity is an appropriate level for most of the days of the period. In December, the simulation model suggests a conservative clinical capacity level, this might be caused by demand seasonality. More historical data is needed to incorporate demand trends and improve the model.

Figure 10. Comparison of suggested clinical capacity and current capacity. Purple line is the actual number of patient seen.

Figure 11 shows the number of surgical cases that can be performed and the number of patients in post-operative care per calendar day. The OR capacity experiences periods of high variation, for example, January has sudden variations from week to week caused by the discretionary time allocated in that month. The variations in OR
capacity have an impact in the census, for example, the number of patients in post-operative care decreases during the last week of December, and it has a sudden increase in January. Census projections could help planning the allocation of post-operative staff for different risk-levels, for example, the orange line could be used for worst-case scenario planning.

![Figure 11](image1.png)

**Figure 11.** Left figure: Estimated number of OR cases performed. Right figure: Estimated census.

The demand for surgery is modeled with the projected number of patients seen in clinic multiplied with the probability 0.68 that a patient will require surgical treatment. Figure 12 shows the estimated number of days for OR access. The average time during the months of October and November is below two weeks. Discretionary days distributed in December have a negative impact on access to surgery, for example, patients seen after January 15 are likely to not be able to access an OR in January. Therefore, major changes are needed in January to increase the surgical capacity of the practice, in particular, within the first two weeks of the month.

![Figure 12](image2.png)

**Figure 12.** Access to OR. Blue line is the average scenario, and red dotted line is the worst case.

**Discussion and conclusions**

Balancing OR and Clinic days in the calendar has made a positive impact on the overall practice. Strategic changes were made to increase the operating room capacity of the practice with limited impact on clinic access. A simulation-based tool was implemented to address some areas of the planning process where improvements could be made. Calendar inefficiencies which may be difficult to identify in a manual process are highlighted by the simulation tool since useful performance metrics such as access to clinic, access to surgery, OR and clinical capacity, and census could be estimated. Most importantly, the simulation tool would help illustrate to the management team the impacts of any changes to the calendar and performed what if scenarios analysis before calendar finalization.

Further improvements are needed to improve the predictive measures of the simulation. The access to OR analysis was done only considering patients accessing OR through clinic, data on transferred patients to the practice will be incorporated to obtain a better model of patient flow. The simulation model currently handles demand seasonality by...
scaling the arrival rates, more HY visit data is needed for a statistical model of seasonality of patent arrival. Finally, an optimization model will be incorporated to determine the best possible allocation of resources.

Acknowledgement

This work is funded in part by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery.

References

Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients.

E. Menti, MS¹, C. Lanera, MS¹, G. Lorenzoni, MS¹, Daniela F. Giachino, MD², Mario De Marchi, MD², Dario Gregori, PhD MA¹, Paola Berchialla, PhD MA³ and Piedmont Study Group on the Genetics of IBD *

¹Unit of Biostatistics, Epidemiology and Public Health, University of Padova, Italy; ²Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino, Italy; ³Medical Statistics Unit, Department of Clinical and Biological Sciences, University of Torino, Italy

*Marco Astegiano, Nicoletta Sapone, Elena Terzi -Gastrohepatology D1 Unit, San Giovanni Battista Hospital, Torino (Director prof Mario Rizzetto);
Angela Sambataro, Paola Salacone, Ezio Gaia - Gastroenterology Unit, San Luigi Hospital, Orbassano;
Rodolfo Rocca, Alessandro Lavagna, Lucia Crocellà, Annalisa Vernetto, Marco Daperno, Angelo Pera - Gastroenterology Unit, Ordine Mauriziano Hospital, Torino;
Silvia Regazzoni, Marco Bardessono - Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino

Abstract

The objective of the study is to assess the predictive performance of three different techniques as classifiers for extra-intestinal manifestations in 152 patients with Crohn’s disease. Naïve Bayes, Bayesian Additive Regression Trees and Bayesian Networks implemented using a Greedy Thick Thinning algorithm for learning dependencies among variables and EM algorithm for learning conditional probabilities associated to each variable are taken into account. Three sets of variables were considered: (i) disease characteristics: presentation, behavior and location (ii) risk factors: age, gender, smoke and familiarity and (iii) genetic polymorphisms of the NOD2, CD14, TNFA, IL12B, and IL1RN genes, whose involvement in Crohn’s disease is known or suspected. Extra-intestinal manifestations occurred in 75 patients. Bayesian Networks achieved accuracy of 82% when considering only clinical factors and 89% when considering also genetic information, outperforming the other techniques. CD14 has a small predicting capability. Adding TNFA, IL12B to the 3020insC NOD2 variant improved the accuracy.

Keywords: Clinical Decision Support, Clinical research informatics, Data mining and statistical data analysis.

Introduction

The extensive clinical heterogeneity of Inflammatory Bowel Disease (IBD), and in particular of Crohn’s disease (CD), has stimulated several efforts to classify patients according to recognized criteria, from the international meetings in Rome (1991) and Vienna (1998) to the recent revision in Montreal [1]. The presence of extra-intestinal manifestation (EIM), among others, has important consequences for the clinical management of CD patients and relevant effects on the overall burden of the disease, the quality of life and the allocation of health resources [2]. Attempts have been made to weight the risk of EIM according to the patients’ conditions both at onset and during disease progression, and among the potential risk factors for onset of EIM the role of several genetic predisposing/modifier factors have been recently reviewed [3], even if their clinical usefulness is at present unclear [4].

This increase of information, usually in conjunction with the limited size of the analyzed samples, is posing several threats to the statistical procedures used for EIM risk stratification. Classical and most used tools, such as logistic
models, are known to have limits in such situations. In addition, the effect of some risk factors on the risk of EIM is known to be non-linear and to interact with other covariates [5, 6].

In a previous paper from Giachino et al. [7], six common statistical models (logistic regression model, generalized additive models, linear and quadratic discriminant analysis, artificial neural networks (ANN) and projection pursuit regression (PPR) were implemented to predict EIM using genetic data in addition of clinical factors, showing the impact that genetics, when appropriately modeled, can have in predicting EIM.

The aim of this paper is to further develop on that pathway, approaching the problem of “predicting” EIM by implementing three different Bayesian classifiers and by assessing its predictive capability in comparison with these, previous and current, results.

Several approaches have been proposed in the Bayesian framework to deal with classification (or prediction) in a clinical setting. Among them, of a heuristically increasing complexity, major groups include naïve Bayes, Bayesian Additive Regression Trees (BART) and Bayesian networks.

Naïve Bayes (NB) classifier applies Bayes’ theorem by assuming that the features are independent given class, regardless of any possible correlation between them. Studied intensively from 1950s, it has been widely adopted in automatic medical diagnosis, with convincing performances, often outperforming other sophisticated techniques, despite its sometime unrealistic independence assumption [8].

Bayesian Additive Regression Trees (BART), having been developed at the beginning for regression problems, is a nonparametric statistical approach making use of a sum-of-trees model and regularization prior on the parameters in order to approximate an unknown function. It has been extended by Chipman et al. [9] to the probit model setup to handle binary classification tasks.

Bayesian Networks (BN) have been introduced in the 1980s as a probabilistic expert system for representing and reasoning models of problems involving uncertainty. Since the beginning of the 1990s, they have been used for developing medical applications [10, 11]. Their success in this field is due to the fact they possess the quality of being both a statistical and an Artificial Intelligent knowledge-representation tool. Furthermore, they allow for structuring domain knowledge by investigating causal relationships among domain variables [12]. In many cases, Bayesian Networks have been proven to outperform other statistical methodology in classification tasks [13].

**Materials and Methods**

The present dataset derives from the larger series of CD and Ulcerative Colitis patients enrolled in our ongoing observational study of IBD genetics in collaboration with three gastroenterology Units in Torino, Italy. An association analysis of the three common NOD2 variants has been reported. Genomic DNA was extracted using a commercial kit (Promega). The nomenclature of the analysed polymorphisms is reported in Table 1, together with references to typing technique and relevant literature. Of the two polymorphisms in the 5' region of the TNFA gene we here consider only the genotype at -308, since all analysed samples were homozygous for the common G allele at the -238 SNP.

**Table 1. Analyzed SNPs.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Polymorphism</th>
<th>Analysis</th>
<th>dbSNP ID</th>
<th>Pr F</th>
<th>Pr R</th>
<th>Restriction enzyme</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOD2</td>
<td>R702W</td>
<td>PCR-RFLP</td>
<td>rs2066844</td>
<td>5’-AGGGAT-GGAGTG-ACATTTC-3’</td>
<td>5’-CGGGAT-GGAGTG-ACATTTC-3’</td>
<td>Msp I</td>
<td>Giachin o et al 2004</td>
</tr>
<tr>
<td></td>
<td>G908R</td>
<td>PCR-RFLP</td>
<td>rs2066845</td>
<td>5’-AGGGAT-GGAGTG-ACATTTC-3’</td>
<td>5’-CGGGAT-GGAGTG-ACATTTC-3’</td>
<td>HhaI</td>
<td>Giachin o et al 2004</td>
</tr>
<tr>
<td></td>
<td>INSC3020</td>
<td>PCR-RFLP</td>
<td>rs2066847</td>
<td>5’-AGGGAT-GGAGTG-ACATTTC-3’</td>
<td>5’-CGGGAT-GGAGTG-ACATTTC-3’</td>
<td>NlaIV</td>
<td>Giachin o et al 2004</td>
</tr>
<tr>
<td>CD14</td>
<td>-159C&gt;T</td>
<td>PCR-RFLP</td>
<td>rs2569190</td>
<td>5’-AGGGAT-GGAGTG-ACATTTC-3’</td>
<td>5’-CGGGAT-GGAGTG-ACATTTC-3’</td>
<td>AvaII</td>
<td>Klein 2002</td>
</tr>
</tbody>
</table>

885
<table>
<thead>
<tr>
<th>Gene</th>
<th>Location</th>
<th>Method</th>
<th>Restriction Enzyme</th>
<th>Restriction Site</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL12B</td>
<td>Ex8 +159A&gt;C</td>
<td>PCR-RFLP</td>
<td>rs3212227</td>
<td>5'-TTTGGAGA-</td>
<td>TaqI C: 161+139bp A: 300bp</td>
<td>D’Alfonso (personal communication)</td>
</tr>
<tr>
<td>IL1RN</td>
<td>86bp VNTR</td>
<td>Electrophoresis</td>
<td>AJ289235</td>
<td>5'-TCTAGGAA-</td>
<td>-</td>
<td>Mansfield et al. 1994</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5'-CACTCCTAT-</td>
<td>-</td>
<td>Vamvakopoulos 2002</td>
</tr>
</tbody>
</table>

**Patients**

We decided to use for this work the same data set as in the previous analysis [7] in order to allow a direct comparison of the various statistical approaches. Detailed clinical and familiar information were acquired from each patient and encoded according to the Vienna classification that was in use at the time of enrolment. Extra-intestinal manifestations were defined as the occurrence of rheumatologic, dermatological, ocular, liver and biliary manifestations and amyloidosis. Patients form a retrospective cohort belonging to the Italian Population. They gave a written consent to the study, which was performed under permission of the Hospital Ethical Committee.

**Naïve Bayes**

A Naïve-Bayes classifier [14] is a simple BN that has the outcome variable as the parent node of all other nodes and no other connections between variables.

Over the BN’s they are easy to construct, since the structure is given a priori and thus no structure learning procedure is needed. They require the assumption that all the features are independent of each other. Despite this strong assumption, Naive-Bayes have proven to outperform many classifiers especially where the features are not strongly correlated [15].

The naïve classifier combines a probability model with a decision rule: it computes the conditional a posteriori probabilities of a categorical class variable given independent predictor variables using the Bayes’ theorem. The metric predictors are supposed to be distributed as a Gaussian. This technique is the simplest class of Bayesian Networks where all of the features are class-conditionally independent. Its simplicity makes it easy to use and it allows to get a good result especially in case of small databases [8]. Moreover, naïve classifiers can be extremely fast in comparison to more sophisticated methods of data mining.

Naïve Bayes classifier’s implementation during this studio takes advantage of the “e1071” R package [16].

**Bayesian Additive Regression Trees**

A Bayesian Additive Regression Trees is a nonparametric Bayesian approach to estimation which uses dimensionally adaptive random basis elements, the regression trees, to approximate an unknown function \( f(x) = E(Y|x) \), by imposing accurately the regularization prior. By weakening the single effects BART ends up with a sum of trees each of which explains a small and different portion of the function \( f \) [9]. Obviously, respect to single trees, models composed of sums of trees have a greater ability to describe into details \( f \), capturing interaction and non-linearity. Hence the information regarding \( f \) is partitioned into different trees, each contributing to the overall fit.

This technique was primarily designed to predict quantitative (continuous) outcomes from observations via regression, but an algorithm that extends BART for binary classification, written in the statistical R package “BayesTree”, is provided online by the original authors of BART [9].
In general, as anticipated, BART model consists of two parts: a sum-of-trees model and a regularization prior of the model’s parameters that keeps the individual tree effects small. Mathematically BART probit extension model for binary classification (coded with outcomes “0” and “1”) can be expressed as:

\[ P[Y = 1|X] = \phi[T_1^M(X) + T_2^M(X) + \cdots + T_m^M(X)] \]

where \( \phi[\cdot] \) denotes the cumulative density function of the standard distribution, \( T \) denotes a binary tree made of a set of node decision rules and a set of terminal nodes, \( M \) a set of parameter values associated with each of the terminal nodes of \( T \). The number of trees, \( m \), is fixed to 200 by choice as a good trade-off: the more this number increases the more the model is flexible, showing excellent prediction capabilities slowing down the computational time.

Through a Bayesian backfitting MCMC algorithm [17] that iteratively constructs and fits successive residuals, thanks to the data augmentation approach of Albert and Chib (1993) [18], the posterior information is extracted. MCMC chooses between different generated trees the one providing the best sum-of-trees model according to a posterior probability. This approach, because of the complex computations, is usually time consuming. However, BART has several appealing features such as the additive characteristic able to catch the variability of the function and the capability to conduct automatic variable selection.

BART approach releases the strong hypothesis of statistical independence of attributes making this technique more akin to real data.

**Bayesian Networks**

A Bayesian Network is a graphical representation of the joint probability distributions over a set of random variables. It consists of a series of nodes representing variables connected by arrows forming a graph that has no cycles. The arcs specify the independence assumptions that must hold between the random variables.

In general, they may be many arcs going into and out of each node, creating a complex network. The most important restriction is that the arcs must not create cycles within the network; the resulting network is known as directed acyclic graph (DAG) [19]. Each node of the network is associated with a set of probability tables. For those nodes without ingoing arcs, the probability distribution is a prior distribution which requires supplying a set of initial values. Both the structure and the numerical parameters of a BN can be learned entirely from data [20, 21].

There are a great number of algorithms for learning the structure and the parameters of Bayesian networks from data. Many of them are based on a scoring function and a search procedure. The algorithms based on a scoring function try to find a graph that best represents the data, according to a specific criterion. They use a scoring function in combination with a search method to measure the goodness of each explored structure from the space of feasible solutions. During the exploring process, the scoring function is applied to evaluate the fitness of each candidate structure to the data.

In this analysis, a variant of this scoring approach is the Greedy Thick Thinning algorithm [22], which optimizes an existing structure by modifying the structure and scoring the result, was performed. By starting from a fully connected DAG and subsequently removing arcs between nodes based on conditional independences tests [23], the Greedy Thick Thinning algorithm is able to isolate the best scoring network. One of the most usual scoring function is the Bayesian metric [24], which is a measure of how likely it is to observe the data given the network structure, i.e. the best network in terms of the Bayesian metric is that one with the highest probability based on the given data [24].

Given the structure of the network, conditional probability learning is done. Since conditional probabilities to be learned depend not just on the parent variables’ values but also on the other linked variable (local structure), usually the assumption each variable is discrete is made. In this way, each local distribution is a collection of multinomial distributions. Given this class of local distributions, probabilities can be efficiently computed when there are no missing data in the sample and assuming local parameter independence, i.e. the probability of each state is independent of the probability of every other state. The learning method performed in this analysis was Expectation-Maximization (EM) algorithm [25].

The EM algorithm performs a number of iterations. For each iteration, the logarithm of the probability of the case data given the current joint probability distribution is computed and the EM-algorithm attempts to maximize this quantity. The starting point of the EM algorithm is the conditional probability tables specified prior to calling the algorithm. As a priori distribution, the uniform distribution was assumed for each variable. The EM algorithm
terminates when the relative difference between the log-likelihood for two successive iterations is sufficiently small (less than $10^{-4}$).

To assess model performance, error rate and predictive value of the Bayesian Network were estimated using a 10-fold cross validation procedure.

Bayesian Networks implementation was carried out using GeNIe 2.0 [26].

**Results**

Basic characteristics of the sample, stratified by occurrence of EIM, are presented in Table 2. A more detailed description of the dataset is given in Giachino’s work [27]. The clinical and genetic characteristics were divided into three groups: (1) characteristics of the disease: age at onset, location, disease behavior and presentation of the disease; (2) known risk factors: sex, smoking behavior and familiarity of the disease; (3) genetic polymorphisms of the NOD2, CD14, IL12B, TNF, IL1RN genes.

**Table 2.** Data description. Median, I, III quartile, number, percentages as appropriate. N indicates the number of cases with a valid information for the given covariate.
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>152</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26.0%</td>
<td>26.7%</td>
<td>26.3%</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>39</td>
<td>36</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50.6%</td>
<td>48%</td>
<td>49.3%</td>
</tr>
<tr>
<td></td>
<td>TT</td>
<td>18</td>
<td>19</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.4%</td>
<td>25.3%</td>
<td>24.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>72</td>
<td>35</td>
<td>53</td>
</tr>
<tr>
<td>TNFA-308</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(45.5%)</td>
<td>(24%)</td>
<td>(34.9%)</td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td>9</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(11.7%)</td>
<td>(5.3%)</td>
<td>(8.6%)</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.5%)</td>
<td>(1.3%)</td>
<td>(3.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>72</td>
<td>49</td>
<td>72</td>
</tr>
<tr>
<td>TNFA-238</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(63.6%)</td>
<td>(30.7%)</td>
<td>(47.4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>72</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>IL12B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(22.1%)</td>
<td>(14.7%)</td>
<td>(18.4%)</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>24</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(31.2%)</td>
<td>(13.3%)</td>
<td>(22.4%)</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>8</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10.4%)</td>
<td>(2.7%)</td>
<td>(6.6%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ILRN*1</td>
<td>72</td>
<td>29</td>
<td>41</td>
</tr>
<tr>
<td>IL1RN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(37.7%)</td>
<td>(16%)</td>
<td>(27%)</td>
</tr>
<tr>
<td></td>
<td>ILRN<em>1/ILRN</em>2</td>
<td>15</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(19.5%)</td>
<td>(9.3%)</td>
<td>(14.5%)</td>
</tr>
<tr>
<td></td>
<td>ILRN*2</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.9%)</td>
<td>(4%)</td>
<td>(3.9%)</td>
</tr>
<tr>
<td></td>
<td>ILRN<em>1/ILRN</em>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.3%)</td>
<td>(1.3%)</td>
<td>(1.3%)</td>
</tr>
<tr>
<td></td>
<td>ILRN<em>2/ILRN</em>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.3%)</td>
<td>(0%)</td>
<td>(0.7%)</td>
</tr>
</tbody>
</table>

Due to the high number of patients with missing values, two different datasets are considered: the first one (1) in which TNF and IL1RN genes were excluded and a second one (2) containing only patients for whom data on TNF and IL were available (Figure 1).

The Bayesian Networks were depicted in Figure 1. Arrows between nodes denoted probability dependencies among variables (bolder arrows pointed out stronger influences among variables).

![Figure 1. Bayesian Network for characterizing EIM: BN1 (left) and BN2 with IL and TNF genes (right).](image_url)

Sensitivity to finding for BN1 is shown in Table 3. As can be seen from the low value of mutual information and variance of beliefs and from the graph, the role of CD14, is negligible for predicting EIM.
Table 3. Sensitivity of finding analysis for Extra intestinal manifestation in BN1.

<table>
<thead>
<tr>
<th>Node</th>
<th>Mutual Information</th>
<th>Variance of beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIM</td>
<td>0.99937</td>
<td>0.2497803</td>
</tr>
<tr>
<td>SMOKER</td>
<td>0.00767</td>
<td>0.0026439</td>
</tr>
<tr>
<td>LOCATION</td>
<td>0.0047</td>
<td>0.0016265</td>
</tr>
<tr>
<td>FAMILIARITY</td>
<td>0.00423</td>
<td>0.0014645</td>
</tr>
<tr>
<td>BEHAVIOUR</td>
<td>0.00338</td>
<td>0.0011712</td>
</tr>
<tr>
<td>PRESENTATION</td>
<td>0.00325</td>
<td>0.0011018</td>
</tr>
<tr>
<td>CD14</td>
<td>0.00095</td>
<td>0.0003287</td>
</tr>
<tr>
<td>AGE</td>
<td>0.00033</td>
<td>0.0001152</td>
</tr>
<tr>
<td>SEX</td>
<td>0.00008</td>
<td>0.0000286</td>
</tr>
<tr>
<td>INSC3020</td>
<td>0.00001</td>
<td>0.0000036</td>
</tr>
<tr>
<td>G908R</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R702W</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In Table 4, comparison of accuracy and predictive values of the three statistical approaches implemented, and of their performance’s enhancements due to the role of genetic variables, was shown.

Table 4. Accuracy, sensitivity, specificity, positive predictive values (PPV) and negative predictive value (NPV) for the three different techniques in case of considering (2) or not (1) the genetic variables.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes1</td>
<td>0.62 (0.54-0.70)</td>
<td>0.57 (0.45-0.69)</td>
<td>0.66 (0.55-0.77)</td>
<td>0.62 (0.50-0.74)</td>
<td>0.61 (0.50-0.72)</td>
<td>0.64 (0.55-0.73)</td>
</tr>
<tr>
<td>Naïve Bayes2</td>
<td>0.79 (0.68-0.88)</td>
<td>0.52 (0.31-0.73)</td>
<td>0.92 (0.80-0.98)</td>
<td>0.75 (0.48-0.93)</td>
<td>0.80 (0.68-0.90)</td>
<td>0.78 (0.67-0.90)</td>
</tr>
<tr>
<td>BART1</td>
<td>0.64 (0.56-0.71)</td>
<td>0.63 (0.51-0.74)</td>
<td>0.65 (0.53-0.75)</td>
<td>0.64 (0.52-0.74)</td>
<td>0.64 (0.52-0.75)</td>
<td>0.71 (0.62-0.79)</td>
</tr>
<tr>
<td>BART2</td>
<td>0.75 (0.63-0.84)</td>
<td>0.26 (0.10-0.48)</td>
<td>0.98 (0.89-1)</td>
<td>0.86 (0.42-1)</td>
<td>0.74 (0.61-0.84)</td>
<td>0.86 (0.78-0.95)</td>
</tr>
<tr>
<td>BN1</td>
<td>0.82 (0.76-0.88)</td>
<td>0.84 (0.77-0.91)</td>
<td>0.79 (0.7-0.88)</td>
<td>0.83 (0.69-0.97)</td>
<td>0.80 (0.69-0.91)</td>
<td>0.85 (0.77-0.93)</td>
</tr>
<tr>
<td>BN2</td>
<td>0.89 (0.81-0.97)</td>
<td>0.78 (0.71-0.86)</td>
<td>0.94 (0.86-1)</td>
<td>0.86 (0.71-0.95)</td>
<td>0.90 (0.81-0.99)</td>
<td>0.95 (0.86-1)</td>
</tr>
</tbody>
</table>

Discussion

The aim of our study was to compare the performance of three Bayesian classifiers in predicting EIM. Among Bayesian classifiers, in this work we focused on NBs, BNs and BARTs, which are the most popular ones. NBs are often regarded as a benchmark and generally well performing models in spite of their simplicity. On the other hand, BNs can be seen as a more general extension of NB, since they can capture also the interaction between features. Finally, BART represents a full Bayesian classifier.
Our analyses showed that BNs outperformed NB and BART. All classifiers show enhancements when introducing the knowledge about IL and TNF genes, paying the price of a small sensitivity. This comparison is useful to understand the important role of this factor in classification tasks. Taking into account the simplicity and the unrealistic assumption of independence at the basis of NB the results obtained through this classifier are quite comparable to those of BART technique that is more complex and computationally heavier. Typically, the performance of Naïve-Bayes can be further improved by carrying out features selection or by relaxing the independence assumption. From a clinical view a potential explanation could be due to the fact the information about genes is very specific. The involvement in Chron Disease of the genetic polymorphism analyzed is not proved but only suspected. Since this disease has a multifactorial origin, the presence of the polymorphism is not necessarily linked to the presence of the extra-intestinal manifestation but the absence can help in excluding it.

This also specifically means that when the information about genes is known the positive predictive values are higher, since better specificity lead to less false positives.

Combining naïve Bayes with features selection is known as selective naïve Bayes [29]. The search strategies for features selection can be carried out following two different approaches: (i) the filter approach and (ii) the wrapper approach [30]. In the filter approach the search strategy is aimed at maximizing the accuracy of the classifier looking only at the discrimination power of the single variables. This is done considering the mutual information function, which is a function independent of the classifier, i.e. the variables already added to the classifier.

Relaxing the independence assumption is basically performed by constructing a tree augmented Naïve-Bayes (TAN), i.e. first learning a structure tree over the set of variables and then adding a link from the response variable to each node, similar to a Naïve-Bayes structure. However, as discussed in [31], NBs often performs well even when the assumption is violated [31].

Finally, NBs rely also on the assumption that continuous variables are Gaussian. The Gaussian assumption means that the conditional probability of each feature given the class is normal with class conditional mean and variance and then it uses maximum likelihood approach to estimate parameters. Since in our data all variables were categorical, the Gaussian assumption was indeed not required.

The present analysis showed that Bayesian Networks were able to provide a further improvement with respect to other statistical model in terms of predictive accuracy. In addition to these features, the graphical nature of BNs allows to display the links between variable. This can facilitate discussion of the model from different backgrounds point of view (clinicians and genetists, for example) and can encourage interdisciplinary. Another important feature of BNs is the ability to learn about the structure and parameters on the basis of observed data. Knowledge of the structure reveals the dependencies of variables and can suggest a direction of causation. However, when using learning algorithm, causal interpretation of dependencies can be a matter of concern and it is more appropriate referring to them as probabilistic relationships.

The assessment of the optimal BN structure is based on the highest probability score for possible candidate structures, given the data provided and eventually penalized for the level of complexity. Different score metrics can be used at this purpose, varying from entropy methods to genetic algorithms. In our analysis we considered a Bayesian metric. The choice of the Bayesian metric as scoring function may lead to have less prediction error compared with BNs suggested by other scoring metrics. However, on the other hand, it may suggest a model that is highly complex and more difficult to interpret with a large number of variables and a large number of links.

A sufficient number of observations is needed to enable a robust estimation of conditional probabilities, even if it has been shown that BNs can yield good prediction accuracy using learning algorithms, even if sample size is small.

As expected for previous studies, our analysis confirms the negligible role of CD14 and that adding the INF-308 and IL12B genotypes information improved the predictive performance of the model. Whereas, adding NOD2, only the variant 3020insC seemed to be associated to predictive accuracy of EIM.

The small sample size and the absence of an independent sample to perform an external validation represent the major limitations of this work, even if the study’s conclusions are strengthened internal validation performed with cross-validation procedure, in order to reduce overfitting bias.

Among non-Bayesian classifiers, Projection Pursuit Regression, Artificial Neural Network and Quadratic Discriminant Analysis outperform in AUC the other models. However, BNs resulted in a comparable accuracy with them. Furthermore, BNs has the advantage to provide an interpretable graphical model, which can be easily discussed and accordingly modified on the basis of medical knowledge of the problem.
In this work we did not focus on the impact of the prior distributions over the posterior probabilities. Usually, the Naïve Bayes approach assumes equi-probable classes as a prior or it uses an estimate for the class probability given by the number of samples in the class over the total number of samples. In absence of prior information from other independent studies, we chose the former strategy in order to use data only once, i.e. just in the training/testing step and not also for specifying the a-priori probability over the classes. Of course as the sample size tend to be large, the prior is forgotten and the data play the most important role is taken by the data.

Also the two BNs were learned in a non-Bayesian way using the K2 greedy search algorithm, which has been shown to outperform other algorithms (Comparison of the Bayesian Network Structure Learning Algorithms).

Regarding BART, [32] proposed a method for incorporating informed prior information about the predictors into the BART model by modifying the prior on the splitting rules as well as the corresponding calculations in the Metropolis-Hastings step. In particular, covariates believed to influence the response can be proposed a priori more often as candidates for splitting rules. Also for this classifier, we chose to use uninformative prior, considering to work in a total ignorance situation.

The presence of small datasets is a common situation in medical applications. BN has some interesting implications for clinical practice where the dataset is usually very small, affecting statistical analysis. BN, even if temporal knowledge is not considered explicitly, can be useful to make a prediction about what will happen in future. In fact the clinical available knowledges about the patients before the treatments is started influence the treatments actions and hence the final outcome. The prognostic Bayesian networks importance in health-care is a well-documented topic and the obtained results in terms of accuracy demonstrates their possible employment is automatic medical prediction.

**Conclusion**

Our study shows that BNs are a feasible and accurate tool for predicting EIM in CD patients. IL and TFN genes influence the classification and bring to a more reliable classification, increasing the accuracy of about 10%.

**References**

Bursting the Information Bubble:  
Identifying Opportunities for Pediatric Patient-Centered Technology

Andrew D Miller, PhD\textsuperscript{1}, Ari H Pollack, MD\textsuperscript{1,2}, Wanda Pratt, PhD\textsuperscript{2}  
\textsuperscript{1}University of Washington, Seattle, WA; \textsuperscript{2}Seattle Children’s Hospital, Seattle, WA

Abstract

Although hospital care is carefully documented and that information is electronically available to clinicians, few information systems exist for patients and their families to use while they are in the hospital. Information often appears trapped within the hospital room. In this paper, we present findings from three participatory design sessions that we conducted with former patients, their parents, and clinicians from a large children’s hospital. Participants discussed challenges they faced getting information while in the hospital, and then designed possible technological solutions. Participants designed technologies aimed at extending parents’ access to and involvement in patients’ care. Their designs showed opportunities for health informatics within and beyond the children’s hospital room: to allow parents and children to disseminate information from within, access information from the hospital room remotely, establish pervasive and collaborative communication with the clinical care team, and learn about their child’s care throughout the hospital stay.

Introduction

The hospital presents a challenging information environment for patients and their caregivers. When patients are active participants in information about their care, readmission rates are lower and errors are less likely to occur (1). Close family members, such as parents, can be essential to patients’ information management practices, helping patients get the information they need and participating in medical decision-making (2). However, getting the latest information about a patient’s condition and care usually requires physical presence in the hospital room waiting for often unpredictable clinical encounters.

Hospitals now have sophisticated information systems that document details of patient care. Yet, few hospitals have systems that provide patients or caregivers with electronic access to that valuable information (3). Electronic medical records (EMRs) are created primarily for clinician documentation, organizing care, and billing, and patient portals are optimized for outpatient needs. Although much information flows to and from patients and their caregivers, hospitals are set up for all that information transfer to occur verbally, within a patient’s room. For example, the primary time for communicating with physicians occurs during the daily rounds, which take place in each patient’s room. Otherwise, patients must rely on sporadic in-person clinician encounters or updates from their bedside nurse. From the patients’ and caregivers’ perspective, information often appears trapped within the hospital room.

In this paper, we present findings from three participatory design sessions that we conducted with former patients, their parents, and clinicians from a large children’s hospital. Participants discussed challenges they faced getting information while in the hospital, and then designed possible technological solutions. Participants designed technologies aimed at extending parents’ access to and involvement in patients’ care, providing more connection between parents, children, and clinicians, and helping parents and children learn throughout their hospital stay. We describe participants’ designs, show how these designs represent their desires for improving patient-centered information systems in the hospital, and explain how those desires translate into opportunities for future design and research.

Related work

Our study builds on related work in the HCI and Health Informatics communities in three areas: inpatient information technologies in the hospital room, technology designs for patients and families, and participatory design as a technique for health and health technology research.

Patient information technologies in the hospital room

In recent years, researchers have designed and evaluated technologies to improve hospital patients’ access to information and enhance their ability to participate in decisions about their care (3). These technologies have the potential to improve upon existing whiteboard and call button-based systems, which while simple to operate can often frustrate patients and easily get out of sync (3,4).
One particular area of focus has been digital displays within the hospital room, such as large wall-mounted digital whiteboards, bedside tablet-based interfaces, or mobile phone apps (5). These technologies have predominantly focused on improving patients’ ability to see information about their care. For example, Wilcox et al. developed a patient-centered large display that could be better kept up to date than manual technologies (6). Dykes et al. built a patient-centered bedside communication center (7) that provides information such as a schedule, test results, and discharge education, and have involved patient preferences in the design of their systems (8). Vawdrey et al. provided a tablet to patients that allowed them to track their progress (9), care plan, and clinical team, and Greysen et al. provided similar information in a mobile phone form factor (10). Kendall et al. show the potential for making background work—‘hidden’ tasks often performed by patients and their caregivers—more visible in the inpatient setting (11). These and other studies have shown the promise for patient-centered technology to improve access to information. However, there is still much untapped potential of information technologies to empower patients as participants and information providers.

**Designs for pediatric patients and families**

The pediatric hospital presents unique design challenges for patient-centered information systems. Patients in this setting have diverse developmental abilities and care decisions are made by or at least in collaboration with their parents or guardians. The primary caregiver (often referred to as informal or family caregiver) (12) often takes an active role in information about the child patient’s care, and in these settings ‘patient-centered’ designs often address parents directly. Chen et al. argue that ‘informal’ caregivers, such as parents, are key stakeholders in decisions about health care and therefore essential users of pediatric health information systems (12). Others have called for the importance of looking at patients and families as whole people, especially in the pediatric setting. Kaziunas et al. studied parent-caregivers of child bone marrow transplant recipients, showing the connections between caregivers’ emotion intertwined with information work and arguing for a focus on ‘transforming lives’ rather than ‘transferring information’ (13). Miller et al. showed the potential for information technology to support caregivers and patients in the inpatient setting, and identified design implications for future patient and parent-centered technologies (14).

Other researchers have worked on designing parent-focused information systems, with most of these studies based in the neonatal intensive care unit (NICU). This setting often requires both intensive parent involvement in care decisions and physical separation from their child (15). As a result, proposed technologies for the NICU involve parents as active participants, both generating information and communicating with health providers through the tools. Safran et al.’s *BabyCareLink*, for example, allows parents to access information about their child from home or work (16). Liu et al. designed a prototype system for parents of high-risk infants that integrated social network features (17). Yet outside of this specialized context, little is known about parents’ and children’s desires and values with respect to information systems that involve them as true participants.

**Participatory design for health**

Participatory design (PD), a technique in which the eventual users of a system co-create ideas and requirements, has been widely used in a variety of settings (18). Using PD, researchers and project leaders can understand the values and priorities of often marginalized or under-represented groups, and the resulting systems better reflect their experiences (18). Health informatics researchers have found PD to be especially useful as a technique for involving patients in the design of hospital technologies. PD has been used in traditional healthcare contexts, such as the design of operating theaters (19). It has also been used to surface the values and priorities of minority populations, such as women from the Caribbean diaspora living with chronic conditions (20) and as part of everyday life (21,22).

Participatory design is particularly useful when designing with and for children (23), and especially when designing health systems. Health-focused human-computer interaction researchers have extended PD techniques for the particular needs of children. Miller et al. worked with young adolescents to design fitness-related games, engaging participants in skits and storytelling (24). Lindberg designed with and for children with cancer (25), finding that designing in pairs was useful, and taking a comics-based approach to help focus on the positive. Others have reflected on the challenges for PD in pediatric healthcare settings, especially for inpatients. Robertson and Balaam questioned the ethics of involving current patients, and suggested working with proxies, such as former patients (26). Researchers have also demonstrated the benefits of involving other representatives, such as family members and clinicians. Bonner et al. co-designed with Child Life Specialists (27). Insights from these projects inspired the design of our study.
Methods

We conducted our study at Seattle Children’s Hospital, a pediatric care hospital in a large urban area in the US, serving patients from a multi-state region. The hospital admits over 15,000 inpatients each year, about half of whom are insured privately. Seattle Children’s Hospital is also a teaching hospital.

Design sessions

We held three participatory design sessions with former patients, parents of former patients, and clinicians at Seattle Children’s Hospital. Sessions were held in meeting rooms at the hospital and lasted between 90 and 120 minutes. The authors’ institutional review board and the hospital approved this work. In Session 1, we recruited only parents and children. In Session 2 and Session 3, both parents and clinicians participated. The same people participated in sessions 2 and 3, which were held one week apart. All sessions were video and audio-recorded; we also took photographs throughout.

In Session 1, parents and children identified barriers to information access in the hospital and worked in pairs to design potential solutions to those problems. The session lasted two hours. We first handed out index cards and asked participants to spend five to ten minutes writing down responses to the prompt “What are some big problems in getting information during a hospital stay?” We then invited participants to share their responses with the group, and summarized these responses on an easel. Participants worked in pairs to pick one of the problems discussed and design a solution for it. We invited participants to make use of a variety of craft supplies in their designs. During this design phase, facilitators circulated to both observe participants’ progress and help them if they were stuck. We then reconvened and each pair shared their design with the group.

Sessions 2 and 3 followed a similar protocol to Session 1, but broken into two 90-minute sessions one week apart, and had both parents and clinicians participating. We were cognizant of the power imbalance between clinicians, parents, and children, and we chose this different format to give parents and clinicians more time to get acquainted, build trust, and discuss issues in depth. In Session 2, we first asked participants to respond to the question “What are some big problems for patients and families in getting information during a hospital stay?” We then held an extended discussion about these challenges.

In Session 3, held one week after Session 2, we began by asking participants to share any relevant issues or thoughts that had occurred to them during the previous week. We then led a discussion of these topics and a review of the issues raised the previous week. We then divided participants into two groups: the two parents worked together, and the three clinicians worked together. We asked participants to pick one problem identified in the discussion and design a technological solution. The groups presented their designs to each other. We then assigned participants to new groups. One parent worked with the physician, and the other parent worked with the nurses. During this second design phase, we invited the new groups to ‘remix’ one of their previous designs, or to work on something new.

Analysis

Our research team met after each session to discuss the sessions and identify potential emergent themes. Once all three sessions were complete, we reviewed the transcripts and design artifacts and iteratively identified themes in the design, checking them against the discussions and making sure the themes were present across both participant groups.

Participants

Thirteen people participated in our sessions: eight parents, two children, and three clinicians. We collaborated with the hospital’s Patient Relations office to recruit parents and children, and recruited clinicians through word of mouth. Parent and child participants were eligible to participate if they had spent at least two days in the hospital within the last two years, felt comfortable engaging in a group discussion in English, and were between the ages of 7 and 75.
Clinicians were eligible if they were currently engaged in full-time patient care within the hospital. Two of the clinicians were nurses, and one was a physician.

**Summary of designs**

Participants in our study created 9 designs, each focusing on a problem identified during the brainstorming discussion. In this section, we provide a summary of each design. In the section that follows, we extract and describe themes across the designs. Designs 1 through 4 were created by parents and children in our first session, without clinicians present. Designs 5 through 9 were created by parents and clinicians during our third session; designs 8 and 9 from this session were inspired by designs 5, 6, and 7. We have given each design a descriptive label (with the exception of design 6, which the participants named themselves).

1. **About me (2 parents)**

These parents sought to address the burden of parents and children repeatedly providing the same information to clinicians and the corresponding danger of forgetting something important in the retelling. They designed a patient-centered website where children and parents can provide information in a centralized manner (see figure 2).

2. **Knowing me, knowing you (2 parents)**

These parents also focused on intake, when parents repeatedly answer similar questions. Their design is a tablet-based interface for capturing patient-generated data, including preferences. This design also includes a list of who has entered the hospital room, including name, photo and role.

3. **Parent-clinician watch (2 children)**

These two former patients focused on the challenge their parents had getting questions answered and connecting with specific clinicians. To solve this problem, they designed a smart watch, which would allow parents to quickly place a video call to a clinician. We asked them why they designed something for their parents and not for themselves. They responded that their parents were the ones who really needed to be active on their behalf, and they worried about children being overwhelmed.

4. **Asking and answering (2 parents)**

These parents focused on challenges parents face both gathering and providing information. Their solution, also tablet-based, allows parents and children to record their answers to common questions, and those answers would travel with the patient when he or she switched services within the hospital. The system would also allow parents to ask questions before or after daily rounds, and promote various ‘hidden’ services overlooked by parents unfamiliar with the hospital, such as volunteers who can look after children while the parent takes a break.

5. **Family-centered portal & communicator (3 clinicians)**

The three clinician participants designed a multi-platform system with three main components: an overview of the care team and the hospital, a patient-focused portal with up-to-date information in a way parents can easily understand, and a communication feature so parents can ask questions and have them answered during clinical team meetings.

---

**Table 1. Participant demographics**

<table>
<thead>
<tr>
<th>Session</th>
<th>Status</th>
<th>Age</th>
<th>Gender</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parent</td>
<td>40-49</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>1</td>
<td>Parent</td>
<td>40-49</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>1</td>
<td>Child</td>
<td>13-17</td>
<td>M</td>
<td>White</td>
</tr>
<tr>
<td>1</td>
<td>Child</td>
<td>13-17</td>
<td>M</td>
<td>Asian/White</td>
</tr>
<tr>
<td>1</td>
<td>Parent</td>
<td>40-49</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>1</td>
<td>Parent</td>
<td>40-49</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>1</td>
<td>Parent</td>
<td>40-49</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>1</td>
<td>Parent</td>
<td>40-49</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>2,3</td>
<td>Parent</td>
<td>30-39</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>2,3</td>
<td>Parent</td>
<td>50-59</td>
<td>M</td>
<td>White</td>
</tr>
<tr>
<td>2,3</td>
<td>Physician</td>
<td>30-39</td>
<td>M</td>
<td>Asian</td>
</tr>
<tr>
<td>2,3</td>
<td>Nurse</td>
<td>18-29</td>
<td>F</td>
<td>White</td>
</tr>
<tr>
<td>2,3</td>
<td>Nurse</td>
<td>30-39</td>
<td>F</td>
<td>White</td>
</tr>
</tbody>
</table>

Figure 2. **About me**, featuring patient-provided video, and medical as well as personal information.
The hospital overview portion gives parents information about amenities, and the name, photo, and role of clinicians who care for their child. The system’s patient-focused portal provides access to the official medical record but in a more interpretable and family-friendly way. This will also include an estimated schedule for the next day, so parents and children can plan ahead. The communication feature seeks to answer the question “what keeps me up at night?” Parents can submit questions 24 hours a day, in the hospital or from home or work, which will be answered by the clinical team and incorporated into the daily rounding activities of the team.

6. OneDocAway (2 parents)

The two parents who participated in the second and third workshops designed two systems. The first of these emerged during a discussion in the second workshop, in which the parents expressed their frustration at feeling ‘trapped’ in the hospital room waiting for clinicians to come by. This system gives an estimated arrival time of the rounding team, showing their progress from patient to patient. They called their solution OneDocAway, based on the transit app OneBusAway, which provides real-time estimates of public transit departure and arrival times.

7. Lab investigator (2 parents)

The parents’ second design addresses challenges understanding lab results and how those results change over time. The system functions as a sort of super-powered patient portal, with results for a particular lab graphed over time, with a high and low marker so parents could know when levels fall outside of normal ranges. Parents could also click on terms for detailed explanations backed by trusted sources vetted by the hospital.

8. Shared calendar (parent & doctor)

In this ‘remixed’ team, one parent and one clinician (a physician) collaborated on a system to improve coordination and collaboration through a parent-focused shared calendar. This system collects events from the medical record, such as medications, labs, rounding times, and specialist visits, and presents a month, week, or day view.

9. Hospital buddy (parent & 2 nurses)

In this second ‘remixed’ team, one parent and two clinicians (both nurses) extended Design 5 to encompass a child and family’s whole experience with the children’s hospital, across inpatient stays and outpatient interactions. The system works on a phone, computer, or hospital-provided tablet. The additions to the system focus on tracking patient/family-clinician communications, with an emphasis on patient and parent-provided information. Parents can also provide information that would persist across shifts and across hospital stays, such as preferences, habits or personality.

Bursting the bubble: opportunities from participants’ designs

Participants’ designs involved parents and children as active participants in the patient’s care—as producers, aggregators, communicators, and learners. They showed how information technology can burst the information bubble of the hospital room, using information and communication technology to move beyond the constraints of the physical room. In doing so, participants designed technologies that allow parents and children to disseminate information to
the care team from within the room, share and receive information while outside the room, stay in persistent contact with their care team, and learn about their condition and care throughout the hospital stay.

Sharing out: parent-initiated dissemination

All participants in our study were adamant in their support of parents and children as participants in their care, and several of the designs encouraged parents and children to take an active role in providing information to clinicians. These designs focused on giving parents and children the ability to pre-emptively share information, so clinicians arrive in the hospital room with a sense of the patient's expertise, priorities, and values.

Many participants spoke about this challenge of sharing out the same information with many clinicians. As one parent put it: “One of the big things…is that that 50 first dates thing that they do, where I think the shortest ER visit we ever had was nine hours. You have just this constant cavalcade of people that come in and ask you the exact same questions.” Clinician participants also recognized this issue; one of the nurses reported it as a very common concern, but said some overlap was inevitable as patients were transferred between specialties. Four of the designs tackle this theme directly, addressing parents’ concerns at being asked the same questions over and over again by allowing them to proactively tell their stories in a way that clinicians would all be able to see. About Me even includes ‘read receipts’ so that parents will know that their messages and updates have been received by specific clinicians. The parents who designed Knowing me, knowing you and Asking and answering both specified that clinicians would ideally view patients’ answers before entering the hospital room, thus arriving informed and ready to ask the important questions. Hospital buddy allows parents to proactively share information such as their child’s personality or routines.

Clinicians in our study were receptive to this kind of approach, although with some caveats. As the doctor in our study put it: “If it’s something critical and important I want to hear it firsthand from the parent. I don’t want to depend on documentation; there’s room for error there. For other things, like, I think family history, social history, other things are important but I’d love if that kind of stuff was better automated and documented and I could say ‘hey I see this’ just ran it real quick by the family and not waste as much time on certain things that potentially are very easy to transfer.”

Designs for ‘sharing out’ also encourage clinicians to regard the patient as a whole person, allowing parents and children to share information that would not be included in standard clinical questions, to share information in a way that makes sense to them but may cut across medical boundaries, and to emphasize certain preferences. For example, About Me includes questions such as “what makes me feel better,” and Knowing Me, Knowing You allows patients to specify preferences, such as an aversion to certain flavors. Those systems also both incorporate parent and child-generated video responses, which add a personal touch by recording information literally in the patient’s own words.

Looking in: accessing the hospital room remotely

Parents, children, and clinicians all designed technologies aimed at extending parents’ access to and involvement in patients’ care, allowing them to keep up to date without trapping them in the patient’s room. Some designs allow parents to virtually peek into the hospital room from afar. In both Knowing me, knowing you and Asking and answering, parents can see a list of who visited the hospital room while they were away. In Family-centered portal & communicator (the clinicians’ design), parents can see a calendar view of visits to the hospital room, as well as a tentative schedule for the upcoming day. The clinicians’ design also incorporates a parent-focused summary of each day, so even remote parents can keep up to date more easily.

Several designs focus on allowing parents to contribute to discussions and care decisions that take place in the hospital room. For example, the Family-centered portal & communicator design allows parents to asynchronously send updates to the clinical care team during clinical rounding. At Seattle Children’s Hospital these rounds are done in or just outside of the hospital room and typically include parents or caregivers who are present at the time of rounding. The Family-centered portal & communicator extends this participation to include parents who cannot be physically present. Clinicians also felt this type of design would help them do their job more effectively. As one nurse put it: “If a parent can’t make rounds then it’s my responsibility as a bedside nurse to relay all that information that happened and I could be prioritizing something different than the parents wanted to hear. And then one provider relaying to another, another nurse relaying something, is difficult.” A technology and ritual in which parents can ‘look in’ to daily rounding even from afar would allow this nurse to focus on other aspects of the patient’s care, secure in the knowledge that the families were well informed and able to provide important information. She would also be able to get a better understanding of how parents were making sense of the information surrounding their child’s care, and the kinds of input they valued.
Other designs focus on giving parents more freedom within the hospital. Several parents spoke about these issues. As one parent put it: “There are many times where I’m like sitting in the hospital room for like hours, like waiting for rounds, although you do kind of know – there is like a four hour time period generally, but that’s a long time, right? And they’re not in your room for very long, so 5-10 minutes, that’s the time that you rush down to grab something to eat because you couldn’t wait any longer.” Another put it more succinctly: “You’re basically held captive in your room unnecessarily.” OneDocAway, which helps estimate the timing of upcoming clinician visits to the hospital room, allows parents to feel confident that they can leave the hospital room and not miss an important conversation with a clinician who stops by the room while they are out.

Connecting across the hospital

Information technology can also improve communication and coordination among parents, children, and clinicians throughout the hospital. Several of the designs enable such pervasive contact throughout the day, allowing parents and clinicians to communicate when they have time, even if both parties cannot meet in the hospital room itself. Parents, children, and clinicians all mentioned the value of transparency and coordinated communication, and several of their designs incorporated ‘anytime’ contact options.

For example, the two child participants in our study designed a smart watch (Parent-clinician watch) for their parents to wear that would support video calls to clinicians, allowing parents to get answers from clinicians without requiring the parents or clinicians to be physically present in the hospital room.

Several participants designed pervasive contact features to ensure their peace of mind at times when clinician encounters were unlikely, especially late at night. Parents in our study seemed keenly aware of the design tension between providing parents with pervasive access to clinicians and overwhelming those clinicians. In the design sessions, parents and clinicians discussed this issue directly:

Parent: “One of the things we started realizing was that you think of something at 11 o’clock at night and you send an email off and then at 11:15 you’re getting an answer back…I learned that I write the email and I save it and I give myself a little reminder that like at eight in the morning or 8:30 I send it to her, because I know she’s going to look at it and I don’t want to disturb her when I shouldn’t be disturbing her.”

Doctor: “That’s very considerate of you to do that.”

Parent: “She’s so amazing.”

The designers of Knowing me, knowing you added an ‘email a clinician’ button. “There’d be a way to email your doctor in the middle of the night (laughs), whether or not they check it. Just sometimes you need to do that, to sleep.”

Other designs focused on another aspect of pervasive contact: proactive awareness. Shared calendar, designed collaboratively by a parent and clinician, allows parents to see a timeline view of their previous and upcoming procedures, labs, and clinician visits, but it also allows parents to add their own schedules to the calendar, which would then be visible by everyone on their clinical team. As the doctor on this team described it, “the parent/caregiver schedule…could automatically go through to the medical team, so the parent just has to say all right, mom’s leaving here, dad’s coming here, then it would automatically be sent over.”

Promoting learning for partnership

Parents can become more effective partners in care as they gain expertise in the medical issues related to their child’s care. Several of the designs reflect this theme by improving parents’ understanding of terminology and helping them interpret test results, so they can better understand and participate in decisions around care.

Parents and clinicians saw the potential of clinical information tools to promote greater understanding. For example, the Lab investigator design allows parents to learn not only whether their child’s values are in the expected range, but lets them drill down into more detailed explanations whenever they wish. As the doctor in our study put it: “It would be nice if the medical records system was built for families and patients to understand as well too. A lot of times we’re saying let me print out the labs for you, explain to you what the labs or the radiology shows but if those labs and the radiology results were written in a way that could be easily understandable and interpretable that’d be nice to have a point of reference for families and patients.”

These tools are especially necessary after an initial diagnosis, when information is flying at parents at high speed while they are trying to cope with the emotional aspects of their child’s condition. As one parent put it: “Education is a very
good point, because we all come in knowing nothing, and we leave with more knowledge than we ever wanted to have but getting from point A to point B can be a long process, a painful process. It would be good to facilitate the learning curve on the early part of it so you’re more participatory. Then you have a certain base level of knowledge that you’re applying to all these results.”

Others specifically noted the potential for technology to help them record clinical information and process it later when they had more time and were better able to cope. “I imagine had there been some sort of video recording or something of some key meetings or things of that sort, where it’s like okay, then I could go back and play what was said by the doctor, that sort of thing, not just relying on my memory, my husband’s memory of what they were saying as I’m sobbing.”

Participants also began to explore the potential for child patients as learners about their own bodies and the teams that care for them. Hospital buddy becomes a totem carried by the child and is used by both the child and parent throughout the hospital experience. The design involves children as participants and learners about their own care, within the inpatient setting and beyond.

**Lessons learned from participatory design with children, parents, and clinicians**

*Parents and clinicians designing together*

When it comes to caring for children, parents and clinicians do not always see eye to eye. In the hospital, parents cede a certain level of control over their child and rely on clinicians’ knowledge and ability. Likewise, clinicians sometimes see parents’ advocacy for their child as a barrier to providing the most effective care. With these issues in mind, we carefully considered the flow and approach of our parent/clinician design sessions. In the end, parents and clinicians worked together quite well. Both sets of participants were accommodating, and frequently expressed their willingness to see things from each others’ perspective. In several exchanges, the two parents in these sessions discussed their worry about overburdening clinicians. Likewise, clinicians frequently expressed their desire for transparency and involving parents actively in decisions about their child’s care.

Much of this considerate attitude is attributable to the participants themselves; both parents and clinicians clearly came to the design sessions ready to collaborate. In addition, we carefully crafted our approach to foster collaboration and a democratic attitude. We began with an icebreaker exercise in which everyone—participants and facilitators alike—created a simple design and shared it with the rest of the group. We did not ask participants to design until the following session, giving them time to build trust and mutual understanding through discussion in their first session together. Only after participants had discussed problems and designed solutions to those problems did we ask parents and clinicians to collaborate. The resulting designs truly were cooperatively designed; parents and clinicians took turns designing and sharing their designs.

Based on our experiences, we are optimistic about the potential for collaborative patient/provider design for the hospital. Through their discussions and designs, parents and clinicians brought out design tensions and learned more about each others’ perspectives, and the designs are all the richer for doing so.

*Sharing video with clinicians*

All the designs created during Session 1 allow parents and/or children to communicate with clinicians via video. In particular, the parents’ designs from this session intend for clinicians to watch the videos before entering the hospital room. As one of the parents who designed *About me* put it, a clinician enters the hospital room “and asks you 25 questions, it gets videotaped and then everyone can watch them before they come in and ask you the same 25 questions!” However, clinical members of our research team quickly pointed out the pragmatic challenges to such an
approach. Clinicians operate in a highly text and graph-driven information workflow with little room to add additional tasks, and the idea of a clinician watching even a brief video before entering each hospital room seems impractical.

What, then, are we to make of parents’ and children’s use of video in their designs? Video-recorded information has definite advantages from a patient’s perspective. Participants felt it would be easier for parents and children to record information by just speaking, which seems like a natural response given that clinicians often ask them to verbally recount medical issues or histories. Low-literacy patients and young children would undoubtedly benefit from a verbal communication tool. Second, allowing patients and their caregivers to share information with clinicians in a naturalistic manner also frees them to express themselves as whole people and communicate across and beyond clinical information categories. Finally, in analyzing the data from our design sessions, we came to understand that parents and children believed because they find video easier to create and consume, that clinicians would also find it easier. These considerations will be crucial to keep in mind in designing interfaces that allow patients to easily express information but also support clinicians’ workflow and need to quickly review text-based information sources.

Limitations and future work

In this study, we worked with parents, children, and clinicians to identify design opportunities for patient-centered information systems in the hospital. However, our study has several limitations. We worked primarily with ‘expert’ parents and children who had extensive experience in the hospital. Despite extensive outreach and recruitment effort, only two children participated in our study. Participants were not demographically representative of the hospital’s patient pool, and Seattle Children’s Hospital is not necessarily representative of all pediatric hospitals. In the future, we hope to verify and extend these findings in other contexts and with different participant groups. Furthermore, participants’ designs significantly extend the ‘status quo’ and challenge existing modes of patient-provider interaction, which may inhibit the successful introduction of such technologies into hospital care. To address this, we plan to design and deploy technology probes based on participants’ designs. Many of the designs are technically feasible, and mainly require canny interaction design and institutional buy-in to be tested in realistic settings.

Conclusion

In this study, we asked participants to answer a simple question: how can technology alleviate big problems patients face in getting information about their care? Through their designs and discussions, participants showed the potential for information technology to allow parents and children to share information to clinicians from within the hospital room; for parents and other loved ones to look in on the hospital room from afar; for parents, children, and clinicians to stay in touch throughout a hospital stay; and for parents and children to learn within the hospital room and beyond. The designs show that patients and their caregivers can be more than recipients of information about their care; they can produce, aggregate, and learn information throughout a hospital stay. The values expressed in these designs will be important to explore in other contexts, such as adult hospitals and chronic condition management as well. While the hospital room will remain the hub for information about a patient’s care, it doesn’t have to be a bubble.

Acknowledgements

We thank our participants, the Patients as Safeguards research team, and Seattle Children’s Hospital. This project was primarily supported by grant #1R01HS022894 from the Agency for Healthcare Research and Quality (AHRQ), with additional funding from the NIH National Library of Medicine Biomedical and Health Informatics Training Grant at the University of Washington, grant #T15LM007442.

References

6. Wilcox L, Morris D, Tan D, Gatewood J. Designing patient-centric information displays for hospitals. CHI


Using Simulations to Improve Electronic Health Record Use, Clinician Training and Patient Safety: Recommendations From A Consensus Conference

Vishnu Mohan, MD, MBI, Deborah Woodcock MBA, Karess McGrath, Gretchen Scholl, BS, Robert Pranaat, Julie W. Doberne, PhD, Dian A. Chase, PhD, MSN, MBA, RN, Jeffrey A. Gold, MD, Joan S. Ash, PhD, MLS, MBA
1Oregon Health & Science University, Portland, Oregon, USA

Abstract

A group of informatics experts in simulation, biomedical informatics, patient safety, medical education, and human factors gathered at Corbett, Oregon on April 30 and May 1, 2015. Their objective: to create a consensus statement on best practices for the use of electronic health record (EHR) simulations in education and training, to improve patient safety, and to outline a strategy for future EHR simulation work. A qualitative approach was utilized to analyze data from the conference and generate recommendations in five major categories: (1) Safety, (2) Education and Training, (3) People and Organizations, (4) Usability and Design, and (5) Sociotechnical Aspects.

Introduction

EHR adoption has increased dramatically in the last decade, primarily as a consequence of the Health Information Technology for Clinical and Economic Health (HITECH) Act of 2009. The widespread implementation of EHRs has resulted in both benefits and unintended consequences, with the latter linked to the potential of causing significant adverse outcomes and patient harm, in part due to poor implementation strategies and EHR customization as well as lack of user training.

There has been a growing realization that improving EHR use may reduce some of these issues, and EHR development, implementation and training have been identified as key areas of research to improve healthcare quality and safety. With this aim in mind, simulation activities, particularly high-fidelity, EHR-specific simulation training, afford an attractive potential solution that directly addresses some of these issues due to the fact that they can create realistic, reproducible environments without any chance of patient risk.

In contrast to a majority of simulation-centric research activities in which the EHR was a tool used by participants as they completed the activity, our prior research has utilized the EHR as the focus of the simulation. We have also developed a model for collaborative intelligent case and simulation design to facilitate EHR training. Having worked on an EHR safety simulation project funded by AHRQ for almost three years, our research team was about to embark on a second project, and the timing was right to gain input from experts regarding the way forward. In order to achieve this, a conference was organized.

Methods

A. Goals of the Conference

The main goals of the conference were to 1) develop a consensus statement on best practices for the use of EHR simulation in education and training, as well as utilizing the EHR to improve patient safety, and 2) outline a strategy for future EHR safety simulation work. The objective was to develop a consensus statement utilizing the input of expert representatives who would bring multiple perspectives to bear, thus providing guidance for practice in the specific area of EHR simulation.

A team of Oregon Health & Science University (OHSU) investigators organized the conference, played a supporting role during the two days when events were ongoing, and subsequently analyzed the data. An experienced facilitator (JA) managed the large group discussions. Small group discussions were led by a volunteer expert (VM, JG, JA) with a research team member assisting.

The agenda was developed so that an international group of experts in simulation, informatics, patient safety, medical education, and human factors (representing each of the five perspectives we hoped to capture) could first assess the state of the art in EHR safety simulation and then recommend a path forward. After reviewing what the
OHSU team had learned during the three years of their research project, the group discussed each of the specific thematic questions detailed in the "Conference Activities" section below from multiple perspectives. Discussions occurred in both large as well as small group formats, and conclusions were reached.

B. Pre-Conference Activities: Planning

Six months prior to the conference, potential participants were identified through literature searches, citation analysis, and by the purposive identification of known experts in the field. Categories of representatives included experts in simulation, critical care clinicians, experts in EHR safety, EHR usability experts and representatives of EHR vendors (see Appendix for a list of participants and their affiliations). Geographic and gender diversity was desired as well as representation from different healthcare systems (university-based healthcare systems, community medical centers, and federal institutions) as well as different EHR types (commercial and home-grown clinical information systems). Many of these attendees were able to represent more than one of the stakeholder groups involved in either EHR usability, ICU management, or simulation.

Participants were provided a number of papers, as well as a copy of the grant used to support the conference, as background material so that a certain level of shared knowledge could be assumed. The plan was to maximize on-site time to enable the sharing of expertise and experiences, generate narratives, and develop consensus statements. Attendees were asked to come prepared to discuss the role simulation could play in understanding and optimizing EHR use and design. In addition, participants were asked to provide insight for future directions and research questions to better develop the role for simulation in clinical informatics.

C. Conference Activities: Discussion

After an initial orientation, participants were tasked with answering the question “what have we learned about simulation for improving EHR safety?” Subsequent to an initial large group discussion on findings and the state of the art, participants broke into three groups, each of which addressed the question “how can we best use simulation to improve EHR safety?” Each group subsequently reported back to the large group.

Next, the group engaged in a large group discussion centered around the theme "what is the role of the EHR in collaborative rounding?", followed by small group discussions to answer the question “how can we best use the EHR for communicative activities (rounding, signoffs, consultations)?” After reporting out to the large group, we held an informal evening reception to expound on the most provocative topics of the day.

The second day began with a large group brainstorming session designed to answer the question “how can we ensure scalability of EHR simulation beyond the ICU to multiple locations and for multiple disciplines?” We ended with a prioritization exercise to identify the most important strategies for moving forward.

D. Post-Conference Activities: Analysis

The research team debriefed at the end of each day of the conference and in the days immediately afterwards. As the qualitative analysis progressed, the team met several times in the months after the conference to discuss insights and construct a consensus of expert considerations.

A total of 1,738 minutes of audio were recorded and transcribed into 25 individual transcripts. Five hundred and sixty-two pages of transcripts were qualitatively analyzed utilizing computer assisted qualitative data analysis software (NVivo, QSR International, Melbourne, Australia). Statements were coded by topic and cluster analysis diagrams employing Jaccard’s coefficient similarity metric were generated to ensure consistent coding between members of the research team. Utilizing a grounded theory approach, the research team identified five major themes: (1) Safety, (2) Education and Training, (3) People and Organizations, (4) Usability and Design, and (5) Sociotechnical Aspects.

Results

Theme: Safety

The expert group addressed issues surrounding EHR safety and the lack of effective tools to address this. One prevailing sentiment was that errors were ubiquitous and universal.

"EHR patient safety technology-based errors, they’re pretty much the same kind of errors around the world.”
It was felt the generation and perception of errors had changed substantially since the advent of the EHR, and that simulation was a prime tool to address safety issues. There is also a paucity of strong epidemiologic data on EHR errors and the true frequency of those reported; the dependence of incident reports probably vastly underestimates the frequency and nature of the spectrum of EHR use errors.

Additionally, significant silos exist within hospital systems between personnel undertaking safety, quality and informatics projects. Simulation can serve as a means to help identify errors in EHR system design and use. Specifically, including informatics expertise in EHR simulation activities could potentially help to identify and correct safety issues, especially those related to errors generated from the use of copy-and-paste techniques, voice recognition, and team-based workflows.

Vendors could also utilize EHR simulations to deliver products that enhance patient safety. Participants felt that the development of specific EHR safety-related standards might provide significant impetus for vendors to incorporate more safety-specific features within their products.

“Vendors want standards because somebody else sets them. It doesn’t favor one vendor over another, and the nice thing is, they don’t have to be responsible for it.”

Experts also emphasized the need to look at more than just the design of the EHR and data structure when identifying potential hazards to patient safety. How the EHR is used and by whom are also critical factors.

“...if you’re talking about HIT safety, you can’t approach this as a technical problem alone.”

**Theme: Education and Training**

The conference emphasized the need to utilize simulation for the purposes of research and education. As the healthcare industry has embraced widespread EHR use, training requirements have shifted away from current models of point-and-click usage to the need for scenario and case-based instruction.

It was also felt that the EHR should be used in simulations in a way that replicates real-world clinical scenarios, with attention to reproducing both realistic workflows as well as the ability to replicate the cognitive complexity of clinical decision-making.

“If the EHR is a tool to help delivery of care, you need to be able to practice using that tool with the functionality you expect it to be used within the real clinical setting. And unless you practice using it that way, you’re never going to know.”

Experts discussed ways to effectively use simulation to improve EHR design as well as optimize the use of the EHR in actual practice. They also discussed utilizing simulation to better train different types of clinicians. This resulted in an examination of the EHR functionality needed to achieve learning objectives.

"There’s a whole series of...strategies how to integrate the EHR into simulations depending on the learning objective. That’s training that I think the vendors could be able to supply to their trainers as well that’s not even part of the normal training."

It was felt that simulation-centric EHR training could be used to identify weaknesses of the system and in user practices, to train for common or rare events, and for formative or summative assessment. Simulations could be used to foster lifelong learning, both in individual and team environments.
"The purpose of simulation is to train people to work together with technology in an environment."

The expert group also appreciated the need to conduct simulations in an interprofessional environment, utilizing levels of complexity that would replicate team-based healthcare delivery, medical student and resident training, and also assist in training other members of the clinical team such as medical assistants, scribes, pharmacists, and nurses.

"...you can't really separate the technology from the environment in which you're training and you can't separate the interprofessional team-based education from the technology use."

The group placed emphasis on the post-simulation debriefing process, and on developing incentives for participation in simulation-based training such as continuing medical education (CME) credits.

It was also felt essential that simulation activities involve clinicians from environments that do not possess the ability to conduct simulations. This could be achieved by using a central simulation center that would act as a resource for multiple organizations and clinicians.

**Theme: People and Organizations**

Participants noted that the current EHR design paradigm still does not match the reality of how EHRs are used by healthcare professionals. Care is delivered by interdisciplinary and interprofessional teams of clinicians; however, EHRs are still designed with the individual user in mind.

“The EHR is designed for a single person when really it should be designed for the team.”

In addition, healthcare professionals serving in different roles use the EHR differently, and often visualize data on EHR screens that display information featuring distinctly different designs.

“...it’s important that simulations can really show how important these different views are, how important the trade-offs between safety and efficiency are.”

While the EHR might be useful in team-based clinical care, its limitations also need to be recognized by end users as well as information system designers.

“The EHR can be a part of the team, but we need to realize what it can and what it cannot also do.”

The experts felt that the EHR might play a crucial role in inculcating a culture of systems-based practice.

“...[the EHR] teaches each member of the team how to interact with each other member of the team and how they can work together to better accomplish the goals than they would individually.”

**Theme: Usability and Design**

EHR system usability and interface design were regarded as key factors in facilitating EHR education and training, and were also seen as crucial elements in ensuring patient safety.
Several factors that have been well described in the literature, such as cognitive overload and user distraction, continue to be a source of EHR-related errors, which suggests that despite attempts to improve EHR interface design and usability, EHR features may still contribute to errors. There is a delicate balance between features which add value and those that distract from completing necessary tasks.

“I will argue that we still have tons of medication errors, even though we have all the alerts that come in the system.”

It was expressed that EHRs need to better support team-based communication and decision-making, and facilitate group decision-making during group rounding.

“…there are a set of tasks that the team assembles to do. And the first tasks have been done by the computer, then you wouldn’t need that team to be there. So it’s all the things that they’re doing that they must converse about that are independent of whatever the computer does. It may be the computer gets smarter and their job gets smaller, but the question is how do you support whatever happens in there.”

While EHR safety features such as alerts are intended to facilitate collaborative clinical care, they can sometimes impede team-based work.

“If an alert comes up and you’re trying to do something, an alert comes up and then I got to deal with that, it stops the conversation.”

Experts grappled with how the user interface should be optimally designed to support education, training, and safety, while still affording for variability in individual users’ performance. One source of inconsistency which may negatively impact patient care is the process of entering data into the EHR. The reliability of data stored within the EHR is affected by the disconnect between the perception of the real-time nature of EHR data and the reality of EHR data entry, which is often delayed due to fragmented clinical workflows. Furthermore, the differences in how each provider uses the data also factors into the overall usability of the system. Incorporating an EHR design model that would account for variability in user tendencies (such as differences in information-finding behavior) could also be a useful EHR feature.

“…each different type of provider uses the data differently, they go to different areas, they look for different things, so we need to build the EHRs so it actually works with whatever the workflow design.”

Participants expressed dissatisfaction over the incongruence between the vendors’ values that guide EHR design (“they’re designing it for efficiency”) and the users’ values that govern EHR usage (“for safety”).

There was extensive discussion about the potential role of simulation in helping guide EHR design to enhance usability. Participants described the potential for developing and deploying a common, standard, simulated patient record that could be EHR-agnostic and contribute to training in different clinical and technological environments.

One major area of unrealized potential for the EHR that was discussed was its role in facilitating handoffs. It was felt that EHR interfaces could be better designed to facilitate provider handoffs that occur not only in person but also remotely (via phone). Clinical decision support tools could be developed to help prioritize the data to convey urgency or clinical severity.
“If we had a decision support tool that helped us synthesize the information better, that’s fantastic. That really helps. If we had a decision support tool that then prioritized it [data]... then all of a sudden you’d have an automatically generated handoff, prioritized handoff note, that actually made sense because it used the same rules to generate itself as the decision process...”

A particular EHR feature that could be helpful to facilitate group communication would be to enable collective situational awareness by sharing EHR screens. This could help others understand what information a user is drawing from to make a decision or to ask a question.

“If you got to design the EHR, it’s got to be designed in the understanding it’s part of a team environment and the person operating the keyboard may not be the person who is actually making the decisions and they might not even be seeing the right screen.”

**Theme: Sociotechnical Aspects**

The expert group felt that the use of EHR simulations in health care was best described within the context of a sociotechnical framework as an interaction among technology, organizations, and individuals. EHR technology is facilitating changes in how clinicians work, and changes in clinician training and work patterns are in turn changing how the EHR is used. We examined this in four contexts: interdisciplinary rounding in an inpatient setting, handoffs, working with a consultant, and providing team-based care in an outpatient setting.

As a consequence of the ubiquitous use of the EHR, the roles of team members have changed while engaging in interdisciplinary rounds.

“...once upon a time the role of the resident was to gather the data, and the attending to provide expert knowledge...Now the resident goes to the EHR where the data is already gathered, and they are doing more of a curation role...The attending’s role is becoming first of all a verification role...Is the right data being gathered and then are they applying it correctly.”

Our experts believed simulation can help develop workflows and processes that balance this shift in clinician roles. The interaction among people in a zero-risk simulation environment can potentially help build interpersonal relationships and assist in the development of trust among clinical care team members. It can also foster systems thinking that allow individuals to better interact with other members of the team and work together to better accomplish group goals.

“...Simulation offers the opportunity I think not only to break down the traditional barriers that exist for data but the traditional barriers that exist for roles.”

There is also a potential for simulation to assist in improving EHR design to better support rounding processes. EHRs have fundamentally changed the way clinicians interact with information. Simulation may help examine this paradigm shift and allow researchers to test new ways that data can be represented to clinicians. Simulation may also help clinicians discover hitherto unidentified errors, and improve safety by preventing patient harm. This requires the acquisition of appropriate tools, both technological as well as cognitive, to facilitate the process.

“This model of simulation training can be extended to a more ideal EHR design with the understanding it is part of a team environment...”
Participants also felt a need to use simulation to test and improve EHRs as part of the collaborative process, especially with respect to consultations, the nature and scope of which have also changed as a consequence of widespread EHR use.

“You don’t physically send your patient. You just simply send a message saying ‘Dr. X, I’ve got this patient I’m concerned about, could you take a look at their chart and give me a recommendation.’ … is it actually a very exhaustive process for the specialist to actually identify their role”

The experts also highlighted new trends that are developing in the use of the EHR to support team-based primary care.

“[In primary care] we don’t round on the patients of the day, we round on 5,000 patients all week. We use population health and reporting tools, but the same paradigm of communication is needed. We need to identify the gaps in care, we need to do something about it, and we need to let the person who is going to carry it out and the patient know about it. So that whole communication tools is needed in different views.”

Interprofessional simulation allows team-based activities to be evaluated, thus promoting improvements at both the individual as well as the group level. It was noted that the EHR was a tool that merely supported clinical workflows and processes, as opposed to defining the activities themselves.

“…decision-making is one team process or function…EHR doesn’t support rounds, it supports the processes and functions…”

Discussion

EHRs were primarily developed to improve workflow efficiency and capture billing in an optimal fashion, and while the explosive increase in EHR implementation has seen unprecedented growth in EHR use within the clinical environment for clinical decision-making, this phenomenon has not been accompanied by an equivalent growth in ensuring patient safety.

“…the vendors are not designing their systems for safety. First, they’re designing them for efficiency and billing, also we’ve got to have a huge paradigm shift that they’re not interested in at all because the people that are buying their systems are chief financial officers, not doctors.”

Recommendation: Simulations should be used to design systems and protocols explicitly to improve usability and patient safety.

Participants clearly felt that EHR safety could be promoted by the use of simulations. Simulations are likely to have a significant salutary effect on team-based care, and utilizing the EHR within simulations in a manner that mirrors real-life clinical scenarios allows the reproduction and modeling of cognitive burden and uncertainty in a way that replicates real-life conditions. Simulations can also assist in optimizing EHR design and functionality; however this advanced use of simulation in health care will also require an equivalent paradigm shift in operational culture.

However EHRs do have the potential to significantly improve safety. In order to achieve this goal, participants felt that a cultural paradigm shift in EHR training was also required.
“So the EHR can facilitate a lot of these[safety-related] things through check-listed tools, but you have to have the appropriate culture change and training to go along with it, as well as a needs assessment, to make sure that you’re creating the right tool.”

**Recommendation: Simulations should be central to clinician training and used to improve clinical team communication.**

Currently simulations are used to some extent in clinician training, but their role is marginal; a paradigm shift that brings simulation front and center in the process of clinician training is also warranted. The emphasis in this new model of training needs to be on the delivery of team-based care, with an equivalent stress on team-based, information-driven decision-making.

“...if you could change the paradigm of the simulation so that it is case-based instead of all of us...in a classroom setting. Just turn it around so that we are all part of a team and attacking the same problem, and that’s our [EHR] training.”

**Recommendation: Simulations should be part of a sociotechnical solution involving a broad spectrum of stakeholders and organizations.**

The effectiveness of the EHR within the context of communication was a theme that arose commonly during deliberations by the participants. Simulation offers the potential of improving communication processes through training and ideas for improved design, although the process will require the involvement of people with multiple skill sets.

“We need to break down these silos, and that includes involving people who understand the many factors, quality, and informatics. Most of the training activities focus on the EHR as a tool, but that doesn’t mean you still don’t need people who understand that tool and [the need to] get the information out.”

The interdisciplinary nature of the expert participants in the conference allowed them to examine potential solutions from an equally interdisciplinary perspective.

In order to achieve this goal, participants believed that it would be necessary to reach beyond the leaders of healthcare systems to engage diverse stakeholders including informaticians, quality improvement groups, EHR vendors, standards organizations, and payors.

“...we have multiple major stakeholders here within this group. We have industry, we have informaticians, we have clinical champions, quality, safety, and education leaders...What does each group need to bring to the table for each of these domains so that we make sure we keep it inclusive, so this doesn’t become just purely something in the value for the informatician or the hospital QI specialist or the hospital trainer?”

**Recommendation: Simulations should be part of a comprehensive incentive program to promote patient safety.**

How can we implement this new model of clinician training? The solution might perhaps require us to define a combination of incentives and penalties - incentives to achieve significant milestones in improving patient safety, and penalties when patient harm cannot be sufficiently eradicated.

“Incentives and penalties work together, you get the political will there, get the leadership to do it, and I think you can do it again and create a market. Vendors aren’t going to do this. If they are going to do it, it’s not because it’s the right thing. They are going to do it for an incentive.”
Of course, any incentive program to promote patient safety utilizing simulations would require the commitment of substantial resources, both to build the technical framework and infrastructure and to develop training and evaluation programs to engage end-users. Since lapses in patient safety are typically associated with costs, both with respect to the care subsequently delivered as well as the costs due to medicolegal repercussions, it might benefit both the risk management and medical malpractice community, as well as health insurance providers to bring resources to bear in order to facilitate the use of simulations to promote patient safety.

“So this...suggestion you made about the medical malpractice companies sponsoring this in some way sounds like the most reasonable approach, and maybe a coalition of vendors and the actual health insurance companies because they’re also going to benefit from fewer use of resources, and fewer costs...I just don’t see the EMR vendors doing this unless somebody’s going to step up and pay for it, you know, and I think that those two groups are going to benefit the most.”

Conclusion

The conference brought together authorities from different disciplines with extensive expertise in simulation, biomedical informatics, patient safety, medical education, and human factors. Their deliberations during the two day meeting were extensive and fruitful. Using a grounded theory approach, data analysis revealed five broad themes that could effectively be used to inform the deployment of simulations to improve clinical education, training, and patient safety.

The themes developed in this paper are intended to act as a framework to develop a comprehensive simulation plan, and assist in decision-making that will place simulation within an appropriate context, and assign a role for simulation-related activities within the infrastructure developed and utilized by organizations to educate and train clinicians and other EHR end-users.

More importantly, the themes also help to underscore the importance of using simulations in today's clinical environment, where the old paradigm of how end users engage with the EHR is rapidly changing not only because of advances in technologies and alterations in clinical workflows, but also notably because of the inexorable advent of the data-driven age of healthcare delivery.

“...how we actually engage with the EHR is antiquated...simulations offer an opportunity to test new modalities of engagement with data...”

References


Acknowledgements

This work was supported by the Agency for Healthcare Research and Quality (AHRQ), grant 1R18HS021637-01. The authors also wish to acknowledge Knewton Sakata, MD for his role in organizing the conference.

Address for Correspondence

Vishnu Mohan, MD, MBI
Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University
Mail Code: BICC
3181 S.W. Sam Jackson Park Rd.
Portland, OR 97239-3098

Appendix

Consensus conference participants, their current affiliations, and their roles: Elizabeth Borycki (University of Victoria, Victoria, BC, nurse, informatician, educator and international perspective), Enrico Coiera (Macquarie University, Sydney, NSW, informatician, educator, and international perspective), John Dulcey (Lansdale, PA, physician and clinical systems consultant), R. Scott Evans (University of Utah, Salt Lake City, UT, informatician and educator), Roy Gill (NextGen Healthcare, Horsham, PA, physician and vendor representative), Richard Holden (Indiana University, Indianapolis, IN, human factors engineer and psychologist), Andre Kushniruk (University of Victoria, Victoria, BC, Canada, informatician, educator, and international perspective), Michael Lieberman (Oregon Health & Science University, Portland, OR, physician and clinician leader), Vincent Liu (Kaiser Permanente Northern California Division of Research, Oakland, CA, physician and research leader), Dean Sittig (University of Texas, Houston, TX, informatician and educator), Adam Wright (Harvard Medical School, Cambridge, MA, informatician and educator), Brian Young (Legacy Health, Portland, OR, physician and clinician leader).

Research team members included: Joan Ash, Jeffrey Gold, Vishnu Mohan, Knewton Sakata, Dian Chase, Julie Doberne, Gretchen Scholl, Deborah Woodcock, Karess McGrath and Robert Pranaat.

Views expressed by participants are their own and not necessarily those of agencies or organizations with which they are affiliated.
Combining Open-domain and Biomedical Knowledge for Topic Recognition in Consumer Health Questions

Yassine Mrabet, PhD¹, Halil Kilicoglu, PhD¹, Kirk Roberts, PhD², Dina Demner-Fushman, MD, PhD¹

¹ Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA
² University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Determining the main topics in consumer health questions is a crucial step in their processing as it allows narrowing the search space to a specific semantic context. In this paper we propose a topic recognition approach based on biomedical and open-domain knowledge bases. In the first step of our method, we recognize named entities in consumer health questions using an unsupervised method that relies on a biomedical knowledge base, UMLS, and an open-domain knowledge base, DBpedia. In the next step, we cast topic recognition as a binary classification problem of deciding whether a named entity is the question topic or not. We evaluated our approach on a dataset from the National Library of Medicine (NLM), introduced in this paper, and another from the Genetic and Rare Disease Information Center (GARD). The combination of knowledge bases outperformed the results obtained by individual knowledge bases by up to 16.5% F1 and achieved state-of-the-art performance. Our results demonstrate that combining open-domain knowledge bases with biomedical knowledge bases can lead to a substantial improvement in understanding user-generated health content.

Introduction

Online resources are increasingly used by consumers to meet their health information needs [1]. According to recent surveys, one of three U.S. adults (35%) looks for information on a medical condition online¹, 30% found online health information helpful for them or someone they know, while only 3% found online information harmful². The same study showed that 15% of Internet users posted questions, comments or information about health-related issues on the web.

The National Library of Medicine (NLM) receives health-related questions from consumers from all over the world. Most consumer health questions received are concerned with disease-related information, such as diagnosis, treatment, and prognosis. At NLM, we have been working towards a system that can automatically answer such questions using available resources. Our previous work focused on intermediate tasks such as question frame extraction [2], question decomposition [3], question type recognition [4], anaphora resolution [2], and spelling correction [5].

Much research in biomedical question answering has focused on answering well-formed clinical questions by professionals [6, 7]. However, consumer health questions differ widely in vocabulary and structure from professional questions. In a recent study, Roberts and Demner-Fushman [8] compared various collections of consumer health questions and professional questions and found that:

- Consumer health questions are closer to open-domain language models while professional questions are closer to medical language models
- Consumer health questions are more focused on medical problems than treatments and tests
- Consumer health questions tend to contain more sub-questions than professional questions.

These characteristics along with high rates of misspellings and ungrammatical sentences pose challenges for automatic question understanding, including tasks such as named entity recognition, and frame extraction. Consumer questions

¹http://www.pewinternet.org/2013/01/15/health-online-2013/
²http://www.pewinternet.org/2011/05/12/social-media-in-context/
about medical problems are often centered on specific diseases, symptoms or treatments, which we refer to as the question topics in this paper. Identifying the topic of a question is a critical step in answering it. Consider the following question:

(1) my question is this: I was born w/a esophagus atresia w/dextrocardia. While the heart hasn’t caused problems, the other has. I get food caught all the time. My question is...is there anything that can fix it cause I can’t eat anything lately without getting it caught. I need help or will starve!.

Note that while two diseases are being mentioned in the question (esophagus atresia and dextrocardia), the topic of the question is only esophagus atresia. While the single disease assumption holds true to a large extent, there are questions with multiple topics, especially when the question is about the relationship between diseases. In the example below, both megalocytic interstitial nephritis and malakoplakia are question topics.

(2) I’d like to learn more about megalocytic interstitial nephritis with malakoplakia.

Question topic recognition can be viewed as an extension of named entity recognition, where the goal is to determine only the entities deemed central to answering the question. In this paper, we propose a novel approach to topic extraction that relies on multi-perspective and knowledge-based recognition of named entities. We use open-domain and biomedical knowledge bases to normalize named entities in consumer health questions. More precisely, we combine disambiguation results from DBpedia[9] and UMLS[10] to provide an initial list of candidate named entities. Starting from this set, we represent the problem of topic recognition as a binary classification task and extend the span of detected topic entities with generic rules. Our experiments show that our method improves the state-of-the-art performance and that the combination of knowledge bases outperforms substantially the individual results, demonstrating that open-domain knowledge bases can benefit information extraction from consumer health text.

Related Work

Duan et al. [11] designed a question analysis approach that tackles the recognition of both the question topic, defined as the main context or constraint of a question (e.g., “Berlin” in “What is the population of Berlin?”) and the question focus, which they introduced as a descriptive feature of the question topic (e.g., “population” in the previous question). Their approach for the identification of question topic and focus is based on MDL (minimum description length) tree cut model. They first represent the question as a chain of base noun phrases, considered as topic terms. Each chain is ordered according to the term specificity in a set of questions and a tree of questions is built from the chains of topic terms. Their cut method then separates the tree into two sets of terms: less specific terms considered as focus and more specific terms considered as topic. While their method allows retrieving multiple topics for one question it requires a rich set of similar questions talking about the same topics, and small sets of questions or dissimilar questions won’t lead to relevant specificity values for the topic terms. While such method could work with open domain questions, it is not adaptable to consumer health questions that tend to be highly heterogeneous in terms of topics, making it difficult to find rich sets of similar questions.

Other methods in open domain used search results to recognize the main entities in user questions. For instance, Carroll et al. [12] proposed a method where questions are first classified as entity triggering questions and then submitted to a search engine. The goal of this process is to find entities associated with the retrieved documents and compare them to the entities mentioned in the question. Candidate entities are then scored according to different factors including the proportion of retrieved documents associated with them or the number of their mentions in the snippets generated from these documents. The candidate entity with the best score is then selected as the main entity (topic) of the question and common features associated with the topic in the retrieved documents are presented to the user. The shortcoming of this method is that it is restricted to questions identified as entity-triggering, which are mostly short open-domain questions. Consumer health questions are more elaborate and often contain several named entities that do not represent the question topic (e.g., disease history or diagnostic tests used to identify the main disease), which makes it difficult to find documents that are relevant to the main topic.
Many studies tackled biomedical question answering [13]. Different approaches have been used for question analysis, ranging from keyword-based search [14] to the extraction of structured frames expressing the relations between entities [15]. Demner-Fushman and Lin[16] proposed an approach to recognize the primary medical problems in clinical questions using UMLS concepts belonging to the Disorder semantic group. This is done by ranking the list of concepts belonging to this category according to their occurrence/position in the analyzed text.

Roberts et al. [3] addressed topic recognition in the scope of the decomposition of consumer health questions. In their approach, candidate named entities were first identified with high recall using a lexicon-based method and then ranked using a support vector machine (SVM) classifier. A series of post-processing rules addressed the term boundaries. The method obtained good results (73.6 and 88.1 F1 score with exact and relaxed match, respectively); however, it did not address multi-topic questions, which represent a quarter of the questions in our dataset.

In general, few approaches tackled specifically the recognition of the main topic in biomedical question analysis. This can be explained by the fact that questions submitted by professionals tend to be concise and tend to have a regular structure which makes the task of topic recognition close to simple keyword extraction. This is not the case for non-expert consumer questions that are more complex to process due to the noise added by numerous peripheral entities.

**Methods**

Our approach to detect topics in consumer health questions is twofold. First, we extract candidate named entities with an unsupervised method. More precisely, we detect and normalize textual mentions with different knowledge bases then we perform named entity recognition using the semantic types associated with the knowledge bases concepts. Second, we classify the recognized entities as a question Topic or not. We call the overall normalization, named entity recognition and entity classification approach $K_{map}$.

**A. Knowledge-based normalization**

We use a concept normalization approach inspired by Mrabet et al. [17]. More precisely, we exploit TF-IDF search and optimization techniques to locate and disambiguate the textual mentions that refer to concepts in a knowledge base $k_i$. This modular process does not require learning corpora, which allows our approach to extend naturally with additional knowledge bases when needed. The implemented algorithm follows the steps outlined below:

1. Split the question into sentences and perform part-of-speech tagging\(^3\).
2. For each sentence, select candidate textual mentions corresponding to a sequence of allowed part-of-speech tags (e.g., nouns, adjectives, adverbs).
3. Use each candidate textual mention as a keyword query to look up concepts in knowledge base $k_i$ based on a TF-IDF search.
4. If no exact match is found between the candidate mention and the knowledge base concepts, the subsequence of words with best TF-IDF score is selected instead. The TF-IDF score is computed by the SolR indexer\(^4\). A minimum threshold is fixed to discard weak matches.
5. For ambiguous mentions (i.e., mentions that have more than one corresponding concept with the maximum TF-IDF score) disambiguation is performed according to global coherence: i.e., we select the knowledge base concept that has more relations with the concepts obtained from other textual mentions. This is done using integer linear programming and contextual search as described in Mrabet et al.[17].

---

\(^3\)We used Stanford Core NLP for POS tagging.

\(^4\)https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html
Table 1: Example type mappings

<table>
<thead>
<tr>
<th>Target types</th>
<th>DBpedia types</th>
<th>UMLS types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>dbo:Disease</td>
<td>Acquired Abnormality</td>
</tr>
<tr>
<td></td>
<td>umbel-rc:AilmentCondition</td>
<td>Cell or Molecular Dysfunction</td>
</tr>
<tr>
<td>Drug</td>
<td>umbel-rc:DrugProduct; dbo:Drug</td>
<td>Clinical Drug</td>
</tr>
<tr>
<td></td>
<td>yago:Medicine103740161</td>
<td>Pharmacologic Substance</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>yago:DiagnosticTest105739043</td>
<td>Diagnostic Procedure</td>
</tr>
</tbody>
</table>

As a result of this process, using a different knowledge base will not necessarily lead to the same candidate textual mentions and will provide different perspectives for the same consumer question.

In the course of this study we are interested in the combination of open-domain knowledge bases and biomedical knowledge bases. We use DBpedia as an open-domain knowledge base [9]. DBpedia is built semi-automatically from Wikipedia infoboxes. For a given Wikipedia article, RDF triples in the format '<subject, predicate, object>' are constructed using the article’s title as subject and the infobox columns as a list of predicates and objects. In a final step, manual mappings were performed by the DBpedia community to normalize predicate names to RDF properties defined in a reference ontology [6]. The 2014 version of DBpedia includes more than 4 million concepts and 3 billion triples.

For the biomedical perspective we use UMLS [10] as a reference knowledge base. We leverage UMLS Semantic Network relations for concept-level disambiguation. In contrast with DBpedia, UMLS relations are not established facts but are defined between semantic types, indicating possible factual relations between the concepts belonging to the given semantic types. For example, the Semantic Network relation Disease or Syndrome-TREATS-Pharmacologic Substance licenses a relation between concepts Cold and Paracetamol. By expanding the subject and object using semantic types hierarchy, we obtain 4,895 relations linking 133 semantic types.

We indexed both knowledge bases with SolR [7] and used LpSolve [8] to solve the disambiguation problem by maximizing the number of relations between the selected concepts.

B. Named entity recognition

We consider all textual mentions successfully normalized in the previous step (i.e., mentions linked to only one knowledge base concept) as candidate named entities. The set of candidate concepts associated with a given mention \( m \), according to a knowledge base \( k_i \) is denoted \( C_{k_i}(m) \). In order to associate a category to a given named entity, we perform a mapping between the semantic types of its associated concept \( c \in C_{k_i} \) and target recognition types (e.g., Disorder, Diagnostic Procedure). This mapping is based on simple regular expressions; for example, we look for the presence of the tokens disorder, disease, syndrome, illness, and pathology to recognize entities referring to medical problems in DBpedia. Table 1 presents some examples of exact mappings from DBpedia and UMLS. In the context of this study, we are primarily interested in the identification of entities referring to medical problems and use the corresponding regular expressions for type mapping.

In order to test the impact of the combination of dissimilar domain point-of-views, we merge the named entities obtained by normalizing the text with different knowledge bases. This merge extends simple union by merging candidate mentions \( m_i, m_j \) that have the same mapped semantic type and overlapping or contiguous text spans into a single mention \( m_{ij} \) such that: \( C_B(m_{ij}) = \bigcup_{i=1}^{N} C_{k_i}(m) \). The text span of \( m_{ij} \) is the text segment covered by \( m_i \) and \( m_j \). The main motivation of this procedure is to obtain higher precision, better text spans and to avoid partial matches.

For instance, in the question “Are there any new research studies enrolling people with carbamoyl phosphate synthase synthase [917]
The two candidate mentions “carbamoyl phosphate synthase” and “synthase 1 deficiency” are merged to form a single mention “carbamoyl phosphate synthase 1 deficiency”.

As we do not use learning corpora, the coverage of our approach is limited to the vocabulary defined in the knowledge bases. To overcome this limitation, we apply a rule-based method to obtain better text spans for the extracted entities as in Roberts et al. [3]. More precisely, we use three rules to extend the span of a named entity with (a) preceding adjectives and modifiers, (b) acronyms occurring directly after the entity and (c) generic keywords occurring directly after the entity (e.g., syndrome, disease, condition).

C. Entity Classification

In the last step of our processing, we approach the task of determining whether an entity is a question topic as a supervised classification task. Given an annotated corpus with gold topic mentions, we build automatically a training corpus for this classification task by considering each recognized named entity that overlaps a gold topic mention as a positive example and each entity with no overlap as a negative example. This translation can be applied to any given annotated corpus and does not restrict the scalability of our approach.

We used a support vector machine (SVM) classifier and experimented with various textual, lexical, and knowledge-based features. After empirical tests we kept only the features that have a minimum Pearson’s correlation factor with the class label. The Pearson’s factor indicates how much two variables are linearly correlated, its values range from -1 (total negative correlation) to 1 (total positive correlation) with 0 indicating no correlation. In our experiments we used a minimum absolute threshold value of 0.1. The first 16 features according to the Pearson’s score are presented in Table 2, along with feature values for the classification of question (3). Disease entities recognized with DBpedia are highlighted, disease entities recognized with UMLS are underlined. In Table 2, candidate named entities are taken from the merge of both results.

(3) Is there any evidence that trauma such as a physical injury i.e., neck injury, torn ligament, etc., can worsen McArdle’s disease?

In our experiments we used the LibLinear9 implementation of support vector machines (SVM).

Evaluation and Discussion

A. Data

To evaluate our method we consider two corpora, one from the Genetic and Rare Disease Information Center (GARD) and the other from the U.S. National Library of Medicine (NLM). The GARD dataset consists of 1459 questions taken from a curated online question set10 with answers created by NIH staff. Each question is associated with a genetic or rare disease (e.g., Beckwith-Wiedemann syndrome, SCOT deficiency, cold agglutinin disease), which is typically the topic of the question. This dataset is publicly available as part of the GARD question decomposition dataset11.

The NLM dataset consists of 263 questions collected from questions submitted to NLM websites (e.g., MedlinePlus, PubMed) and manually classified as consumer health questions with at least one disease topic. The question writers are generally concerned with a particular medical problem, though not always a named disease (e.g., pain in right ear, difficulty walking), and are asking for general medical information (such as that available on MedlinePlus), examples (1), (2) and (3) were taken from the NLM dataset. The dataset contains identifying information and is thus not publicly available. A de-identified set of questions largely overlapping with this dataset is in preparation for public release. Table 3 presents more details on each corpus.

9https://www.csie.ntu.edu.tw/~cjlin/liblinear/
10https://rarediseases.info.nih.gov/gard
Table 2: Main topic classification features for question 3.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>trauma</th>
<th>physical injury</th>
<th>neck injury</th>
<th>McArdle’s disease</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ST</strong></td>
<td>UMLS semantic type (for mentions linked to UMLS entities).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TFIDF1</strong></td>
<td>TF-IDF of entity in DBpedia.</td>
<td>11.63</td>
<td>18.33</td>
<td>11.38</td>
<td>9.33</td>
</tr>
<tr>
<td><strong>TFIDF2</strong></td>
<td>TF-IDF of entity in UMLS.</td>
<td>6.99</td>
<td>14.03</td>
<td>12.98</td>
<td>10.31</td>
</tr>
<tr>
<td><strong>Ambiguity</strong></td>
<td>relative number of entities sharing the maximum TF-IDF score from a target knowledge base. Average ambiguity is considered if an entity was detected by both DBpedia and UMLS.</td>
<td>25.65</td>
<td>0.00</td>
<td>12.50</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>P − C1</strong></td>
<td>Relative character position of first occurrence.</td>
<td>0.21</td>
<td>0.34</td>
<td>0.51</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>P − C2</strong></td>
<td>Relative character position of last occurrence.</td>
<td>0.21</td>
<td>0.34</td>
<td>0.51</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>P − E</strong></td>
<td>Position in the set of entities.</td>
<td>0.25</td>
<td>0.5</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Freq.</strong></td>
<td>Entity frequency in the question.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>CI</strong></td>
<td>Character length</td>
<td>6</td>
<td>15</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>Total nb. of entities in the question</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>Entity tokens number</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>W_1</strong></td>
<td>Entity token 1 (W_1).</td>
<td>trauma</td>
<td>physical</td>
<td>neck</td>
<td>McArdle</td>
</tr>
<tr>
<td><strong>L_1</strong></td>
<td>Lemma of W_1.</td>
<td>trauma</td>
<td>physical</td>
<td>neck</td>
<td>McArdle</td>
</tr>
<tr>
<td><strong>POS_1</strong></td>
<td>Part-of-speech tag of w_1.</td>
<td>NNS</td>
<td>JJ</td>
<td>NNS</td>
<td>NNS</td>
</tr>
<tr>
<td><strong>L_{W−1}</strong></td>
<td>Lemma of token preceding the entity (W_{W−1}).</td>
<td>that</td>
<td>a</td>
<td>,</td>
<td>worsen</td>
</tr>
<tr>
<td><strong>POS_{W−1}</strong></td>
<td>Part-of-speech tag of W_{W−1}.</td>
<td>IN</td>
<td>DT</td>
<td>,</td>
<td>VB</td>
</tr>
</tbody>
</table>

Table 3: Corpus Statistics

<table>
<thead>
<tr>
<th>Features</th>
<th>GARD</th>
<th>NLM Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions</td>
<td>1459</td>
<td>263</td>
</tr>
<tr>
<td>Multi-topic questions</td>
<td>46 (3.15%)</td>
<td>64 (24.33%)</td>
</tr>
<tr>
<td>Topic per Question</td>
<td>1.03</td>
<td>1.25</td>
</tr>
<tr>
<td>Tokens</td>
<td>46,964</td>
<td>17,848</td>
</tr>
<tr>
<td>Tokens per Question</td>
<td>32.18</td>
<td>67.86</td>
</tr>
</tbody>
</table>

On the NLM dataset, our NER process retrieved an average of 10.61 entities per question with DBpedia and 9.68 with UMLS. On the GARD dataset, our NER module retrieved 6.59 entities per question with DBpedia and 6.33 with UMLS.

B. Topic Recognition Results

To test our approach we perform a 10-fold cross validation to associate a class label (topic/none) to each candidate entity (the label is assigned to the entity when the latter is part of the 10% test split). Table 4 presents the results for topic recognition evaluated with 10-fold cross-validation on each corpus.

The merger of both knowledge bases performed substantially better than the individual knowledge bases (cf. Table 4) with an increase of 16.6% in F1 score on the NLM dataset and 2.1% on the GARD dataset for partial matches. For exact matches, the increase is of 3.5% on the GARD dataset and 10.6% on the NLM dataset. Besides the increased recall, this result can also be explained by the fact that our merge method is able to extend the spans of textual mentions from two annotation points-of-view; i.e., the open-domain perspective brought by DBpedia and the medical perspective from UMLS. The number of questions with multiple topics in the GARD dataset is not significant enough to draw reliable conclusions (only 3% of all questions), however we can observe that recall on these questions is above 70%. The NLM dataset contains more multi-topic questions (24.3%) and we can observe that recall on these questions
is more than 10% better than the overall recall with a 12% improvement in precision for exact match, which shows that our approach has higher effectiveness on these questions. This may be explained by the fact that consumers who write multiple-topic questions are looking for a more-precise and fine-grained information requiring a more focused writing to express the link between the different topics, which eases the named entity recognition and classification process.

In Table 5, we compare our results on the GARD dataset with the state-of-the-art approach, denoted \( QD[3] \), which uses a support vector machine over a set of semantic and word-level features to rank an initial set of candidate mentions provided by a lexicon constructed from UMLS. This approach also uses a set of rules (\( R_1 \)) to extend the spans of the detected topic mentions (e.g., appending generic words such as “disease” or “syndrome” if they follow the detected entity in the text). The comparison should be taken with caution as we use a different set of rules to extend entity boundaries. In order to reach more accurate observations, we use both a subset of \( R_1 \) that can be applied to our system, denoted \( R'_1 \) and our full set of rules, denoted \( R_2 \). The \( K_{map} + R_2 \) results correspond to the full-method results presented in Table 4.

Table 5: Comparison with state-of-the-art approach on the GARD dataset.

<table>
<thead>
<tr>
<th>Approach</th>
<th>GARD</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact Match</td>
<td>Partial Match</td>
<td>Exact Match</td>
<td>Partial Match</td>
<td>Exact Match</td>
<td>Partial Match</td>
<td>Exact Match</td>
</tr>
<tr>
<td>( QD )</td>
<td>56.4</td>
<td>54.7</td>
<td>55.6</td>
<td>89.2</td>
<td>86.5</td>
<td>87.9</td>
<td></td>
</tr>
<tr>
<td>( QD + R_1 )</td>
<td>74.8</td>
<td>72.5</td>
<td>73.6</td>
<td>89.5</td>
<td>86.8</td>
<td>88.1</td>
<td></td>
</tr>
<tr>
<td>( K_{map} )</td>
<td>70.0</td>
<td>68.3</td>
<td>69.1</td>
<td>92.4</td>
<td>84.9</td>
<td>88.5</td>
<td></td>
</tr>
<tr>
<td>( K_{map} + R'_1 )</td>
<td>75.3</td>
<td>76.5</td>
<td>75.9</td>
<td>91.0</td>
<td>90.5</td>
<td>90.8</td>
<td></td>
</tr>
<tr>
<td>( K_{map} + R_2 )</td>
<td>80.4</td>
<td>81.7</td>
<td>81.0</td>
<td>91.4</td>
<td>90.5</td>
<td>90.9</td>
<td></td>
</tr>
</tbody>
</table>

Using a the subset of rules \( R'_1 \) improved performance and the final set of rules \( R_2 \) led to an overall performance increase on both exact matching (+7.4% F1) and partial matching (+2.8% F1) over \( QD + R_1 \). The overall improvement with respect to \( QD \) can be explained by several factors:

- \( QD \) does not tackle multi-topic questions and extracts only one topic for a given question, which reduces recall. More generally, ranking-based methods do not provide a systematic solution to multi-topic questions. Even if a threshold can be fixed to get the \( N \) best entities, relevant extensions are needed in order to determine the number of topics in a question. Binary classification solves this problem as it processes each entity individually from both a local and contextual point of view regardless of the rank/importance of other entities in the question, which provides a natural way of selecting the number of entities to be considered as topic/important in a question.

- Vocabulary coverage: this factor is hard to estimate exactly. While our method used DBpedia as an open-domain knowledge base, it also did not have the same level of coverage on the UMLS part, as it annotates only nouns or noun phrases that correspond to UMLS concepts, while \( QD \) is based on an extended lexicon that is able to detect individual keywords and to annotate verbs and adjectives.
Prior disambiguation: in $K_{map}$ entities are disambiguated using knowledge base relations while $QD$ uses only a lexicon-based method, which helps provide a more reliable set of candidate entities to the binary classifier.

We analyzed two random sets of 20 errors, respectively from the GARD dataset and from the NLM dataset on the basis of exact match evaluation. Table 6 presents the observed error types with percentages and examples.

Table 6: Error Types on Random subsets of GARD and NLM datasets

<table>
<thead>
<tr>
<th>Error type</th>
<th>Error %</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Normalization</td>
<td>15%</td>
<td>Some acronyms like APL and ARVC were not disambiguated.</td>
</tr>
<tr>
<td>Entity not in the knowledge bases</td>
<td>15%</td>
<td>The system recognized only agonizing disease instead of this very painful and agonizing disease, or two foci: hereditary neuropathy and liability to pressure palsy instead of one single hereditary neuropathy with liability to pressure palsy.</td>
</tr>
<tr>
<td>Classification</td>
<td>45%</td>
<td>autism not classified as Topic in two questions.</td>
</tr>
<tr>
<td>Expansion Rules</td>
<td>15%</td>
<td>multiple cerebral cavernous malformations (CCM) instead of cerebral cavernous malformations (CCM).</td>
</tr>
<tr>
<td>Misspelling</td>
<td>5%</td>
<td>The question mentions Mycobacterium fortuitum instead of Mycobacterium fortitum.</td>
</tr>
<tr>
<td>Benchmark Error</td>
<td>5%</td>
<td>this disease annotated as Topic instead of the two question topics hypokalemic periodic paralysis and hyperkalemic periodic paralysis</td>
</tr>
<tr>
<td>Entity not in the knowledge bases</td>
<td>25%</td>
<td>sounding in my ear, liver is damaged, pain in my shoulder, growth on the neck.</td>
</tr>
<tr>
<td>Classification</td>
<td>60%</td>
<td>strokes, heart attacks, Nph, shingles recognized as entities but not as Topic in some questions. TX wrongly classified as Topic.</td>
</tr>
<tr>
<td>Misspelling</td>
<td>15%</td>
<td>canker, Thalassamia, hitiala hernia.</td>
</tr>
</tbody>
</table>

Classification was the main cause of errors on both datasets, with a more noticeable impact on the NLM dataset, which is clearly harder to process due to the lack of training data (263 questions vs. 1459 in the GARD dataset) and to longer queries (NLM requests are two times longer than GARD requests in average, cf. Table 3). Other errors vary by dataset. In GARD, entity normalization was the second source of errors (15%) as some acronyms and expressions could not be disambiguated with our knowledge-based approach. At a similar rate, the lack of coverage (concepts) in the knowledge bases did not allow detecting some topics. The impact of coverage seemed to be more pronounced on the NLM dataset, where spatial expressions play an important role in problem descriptions (e.g., pain in my shoulder). Errors due to misspellings were also more frequent in the NLM dataset, which may be explained by the fact that these questions are not edited for spelling, as it was done for the GARD dataset.

C. Ablation study

In this section, we study the classification of every instance of recognized entities, as opposed to finding an overall topic of the question as described above. This internal evaluation is important as achieving high classification accuracy is required to assess the scalability of our approach in distinguishing topic entities from other entities. We also tested our classification method with each knowledge base individually, and then with the merger of the knowledge bases results. We report ROC Area and precision, recall and F1 score for the Topic class in Table 7.

Our classifier reached high ROC area on the GARD corpus (93.0%) but obtained a lower performance on the NLM consumer health questions dataset (76.2%). This is partly due to additional noisy entities in the NLM questions which are generally longer than GARD requests, and thus have a lower topic-to-named-entity ratio. NLM questions are also less grammatical and don’t target genetic/rare diseases, which makes the topic term less likely to be a typical
Table 7: Evaluation Results for Binary Classification (10-fold cross-validation).

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>GARD</th>
<th>NLM Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROC Area</td>
<td>P</td>
</tr>
<tr>
<td>DBpedia</td>
<td>92.6</td>
<td>92.8</td>
</tr>
<tr>
<td>UMLS</td>
<td>91.9</td>
<td><strong>93.0</strong></td>
</tr>
<tr>
<td>Merge</td>
<td><strong>93.0</strong></td>
<td>92.6</td>
</tr>
</tbody>
</table>

named entity (e.g., “loss of vision” vs. “streiff syndrome”). This makes it more difficult to detect Topic entities with position-based features and TF-IDF scores in the NLM dataset.

The combination of both knowledge bases obtained the best classification performance on both corpora. While higher recall was anticipated, higher precision can be partly explained by the merge heuristic that is performed at entity level which leads to better text spans. The different perspectives of the knowledge bases may also allow the SVM to reach more accurate hyperplanes.

To better understand the behavior of the classifier, we also studied the role of the main corpus features. Table 8 presents the Pearson’s correlation score of the best features for both datasets.

Table 8: Pearson’s correlation score for the best corpus features

<table>
<thead>
<tr>
<th>Corpus Features</th>
<th>GARD</th>
<th>NLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character Position</td>
<td>.58</td>
<td>.31</td>
</tr>
<tr>
<td>Entity Position</td>
<td>.54</td>
<td>.37</td>
</tr>
<tr>
<td>Character Length</td>
<td>.45</td>
<td>.22</td>
</tr>
<tr>
<td>Ambiguity</td>
<td>.42</td>
<td>.15</td>
</tr>
<tr>
<td>Entity Nb.</td>
<td>.42</td>
<td>.27</td>
</tr>
<tr>
<td>Entity Token Nb.</td>
<td>.35</td>
<td>.15</td>
</tr>
<tr>
<td>Entity Word 1 POS</td>
<td>.32</td>
<td>.18</td>
</tr>
<tr>
<td>UMLS Semantic Type</td>
<td>.24</td>
<td>.10</td>
</tr>
</tbody>
</table>

Topic entities often occur at the beginning of consumer health questions which explains the impact of position-related entities on both datasets. Ambiguity was interestingly among the best features. High ambiguity values indicate that an entity is an overly general (open-)domain entity. If such entity is associated with the medical problem type it would be a relevant indicator for positive classification (like generic disease terms such as syndrome), if it is not associated with medical problems it can be a relevant indicator for negative examples. If we compare the features ranking between the two corpora, we note that the semantic type information is substantially more correlated with the class label in the GARD dataset, which suggests a more heterogeneous set of semantic types for topics in the NLM dataset. Ambiguity is also more highly correlated in the GARD dataset, which suggest a more regular use of generic disease terms. More generally, the tokens and lemmas are not among the best features which suggests that our approach may also be ported to other corpora as a baseline for topic recognition in corpora from other domains.

Conclusions

We presented a novel approach to recognition of topics in consumer health questions using both open-domain and biomedical knowledge bases. Our experiments showed that the combination of such knowledge bases leads to substantial improvement. Our binary classification approach was also able to tackle effectively multi-topic questions. Comparison on a standard benchmark also showed that our classification approach outperforms an SVM-based ranking approach. In future work, we plan to include additional knowledge bases to improve the recall of our method by taking into account spatial relations. We will also explore other semantic features such as WordNet synsets to improve the contextual knowledge for entity classification.
Acknowledgments

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine and by an appointment of Y. Mrabet to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. K. Roberts was supported by NLM grant 1K99LM012104.

References

Using Drug Similarities for Discovery of Possible Adverse Reactions

Emir Muñoz, MEng\textsuperscript{1,2}, Vít Nováček, PhD\textsuperscript{2}, Pierre-Yves Vandenbussche, PhD\textsuperscript{1}
\textsuperscript{1}Fujitsu Ireland Ltd., Co. Dublin, Ireland;\textsuperscript{2}Insight Centre for Data Analytics at NUI Galway, Co. Galway, Ireland

Abstract

We propose a new computational method for discovery of possible adverse drug reactions. The method consists of two key steps. First we use openly available resources to semi-automatically compile a consolidated data set describing drugs and their features (e.g., chemical structure, related targets, indications or known adverse reaction). The data set is represented as a graph, which allows for definition of graph-based similarity metrics. The metrics can then be used for propagating known adverse reactions between similar drugs, which leads to weighted (i.e., ranked) predictions of previously unknown links between drugs and their possible side effects. We implemented the proposed method in the form of a software prototype and evaluated our approach by discarding known drug-side effect links from our data and checking whether our prototype is able to re-discover them. As this is an evaluation methodology used by several recent state of the art approaches, we could compare our results with them. Our approach scored best in all widely used metrics like precision, recall or the ratio of relevant predictions present among the top ranked results. The improvement was as much as 125.79% over the next best approach. For instance, the F1 score was 0.5606 (66.35% better than the next best method). Most importantly, in 95.32% of cases, the top five results contain at least one, but typically three correctly predicted side effect (36.05% better than the second best approach).

Introduction

Adverse Drug Reactions (ADRs)\textsuperscript{a} can severely limit the intended benefit of drugs and accounts for a large number of hospital admissions, 42% of which could be prevented\textsuperscript{1,2}. ADRs can result in reduction of the patients’ quality of life or even death in extreme cases\textsuperscript{3}. The use of machine learning techniques has now become a common practice to improve drug safety and in particular to detect ADRs. However, many of the state of the art side effect detection systems and procedures depend on patient records or explicit incident reports\textsuperscript{3--5} and therefore assume ADRs already demonstrated within a population. Stakeholders in the drug development and administration lifecycle could greatly benefit from a technique that would help palliating drug’s ADRs before it is released on the market. The presented work addresses the area of computational side effect discovery using information in openly available biomedical databases. In recent years, increasing volume of biomedical data has been openly published online. This includes structured resources like Drugbank\textsuperscript{6} and SIDER\textsuperscript{7} that are represented in a machine-readable and interchangeable standard format in the Bio2RDF project\textsuperscript{8}. The uniform format allows for easy combination of these resources that can help to get different viewpoints on biomedical facts within one interlinked resource --- knowledge graph, an increasingly used umbrella term for loosely structured graph-based knowledge representation\textsuperscript{9}. Knowledge graphs are well suited to discovery of implicit knowledge hidden in the data, which we utilise in our approach to discovery of adverse drug reactions. We infer new links between drugs and side effects using side effect propagation along drug similarity relationships computed using the contents of consolidated Bio2RDF graphs.

The key contributions of our approach can be summarized as follows: (1) Best results in seven metrics traditionally used in the field (e.g., precision, recall or number of correct predictions among the top-ranked results) when compared to recent related works\textsuperscript{10--14}. (2) Best performance in being able to discover actual side effects and rank them so that they appear at the top of the results. In 95.32% of cases, the top 5 results contain at least one correctly predicted side effect (36.05% better than the second best approach). Moreover, the top 5 results typically contain at least 3 correctly predicted side effects. For drug of certain types (like NSAIDs or barbiturates), all the top 5 results are typically correct discoveries\textsuperscript{b}. (3) Superior flexibility - our prototype is able to incorporate many relevant data sets automatically, while all other related approaches would need to either develop additional ad hoc pre-processing tools, or replicate the presented technology.

\textsuperscript{a}Note that we also use several synonyms of ADRs like “adverse drug events” or “drug side effects” interchangeably in the paper.\textsuperscript{b}This is a very important aspect of the presented method, since in many practical applications, it is crucial to provide high-quality results among the top few ones. To give two examples: In clinical applications, physicians have very little time and therefore any automated predictions the use for their decisions have to be concise and highly reliable. In pharmaceutical research, the computational discoveries have to be tested in expensive and long laboratory experiments, and thus a more concise and reliable method like ours can potentially save money and time.
The expected benefits of the presented technology for its target audiences are:

- fully automated prediction of possible side effects;
- flexibility due to support for semi-automated incorporation of new training data sets;
- applicability to decision support in
  - clinical practice -- offering instant, reliable and concise feedback on possible side effects of drugs in daily use, helping to save lives and prevent complications in patients;
  - pharmacological research -- providing comprehensive lists of potential side effects of new compounds based in their similarity to existing drugs, helping to develop drugs faster and save money, while contributing to their safety.

A current limitation of the presented technology is its ability to make predictions about slightly smaller number of drugs than some related approaches. Still, this does not prevent from practical applications bringing the above-mentioned benefits, as explained in detail in the discussion section.

System Overview

The drug-side effect link prediction system consists of two phases: P1) an offline phase where the data processing takes place (integration of the heterogeneous data sources to build a knowledge graph in the RDF format\cite{15}, which is then used for computation of drug similarities), and P2) an on-line phase that provides interfaces for users to query the drug similarities database. Each phase consist of different module interactions to either build the similarities database, or query it. Figure 1 shows the system architecture and components involved in both phases.

![Figure 1. System architecture.](image)

Methods

We gathered all data sets into the pivot RDF format. The formal specification of RDF is convenient, as it allows for the integration of data while preserving their original semantics. We use a graph representation to model the underlying drug-related background knowledge.
**Data sources**

The main data sets we used are DrugBank for drugs, SIDER for drug side-effects, and PubChem for compound IDs which are used to link drugs in DrugBank to the ones in SIDER. Although all the mentioned sources have public access dumps, our method takes advantage of the graph representation using RDF, therefore we use the transformed data made available in the Bio2RDF project\(^8\) (release 4, accessed in December, 2015).

<table>
<thead>
<tr>
<th>Data set</th>
<th>Type</th>
<th>Source and version</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrugBank</td>
<td>Database</td>
<td>Bio2RDF v2015-12-06</td>
<td>Drug types, chemical information</td>
</tr>
<tr>
<td>SIDER</td>
<td>Database</td>
<td>Bio2RDF v2015-12-06</td>
<td>Side effects of drugs</td>
</tr>
</tbody>
</table>

The RDF graphs with all data sets are stored in an Apache Jena Fuseki\(^c\) triplestore and accessed using the W3C recommendation SPARQL 1.1\(^d\) query language for RDF.

**Data processing**

The first step in the data processing pipeline is indexing the Bio2RDF data sets to facilitate a uniform access to them. As shown in Figure 1, the data sources considered herein are drug profiles and side effects, which are integrated by the module data integration, to later make them available for querying as an RDF graph using HTTP SPARQL requests. Even when we use different public sources to build the RDF knowledge graph, by design our method is able to work with other sources such as post-market reports. Because of the flexibility of RDF, other structured sources can be used with information about drugs (e.g., chemical, biological, phenotypic), side-effects, drug classification, and diseases.

The number of unique approved small-molecule drugs gathered from the structured data sets was 731, while the number of side-effects gathered was 4,652. Finally, the resulting RDF knowledge graph contains a total number of 10.7 million unique statements.

**Measures**

*Resource Features Vector.* Given a resource (e.g., a drug node in our specific case), we extract a set of features that represent the connections between the resource and other resources in the graph. Formally, we extract the set of incoming and outgoing relationships of a resource \(X\) by using query patterns \(((?,X), (?,?)\) and \((X,?)\), respectively. For example, let \(A\) be the set of features for resource \(a\) in Figure 2. Node \(a\) in the RDF knowledge graph has outgoing relations to the resources \(c\), \(e\), and \(f\) with predicates \(\ell_1\), \(\ell_3\), and \(\ell_4\), respectively; and a unique incoming relation from resource \(d\) with predicate \(\ell_2\). Then \(A\) is defined as follows:

\[
A = \text{Features}_{LD}(a) = \{(\ell_1, c), (\ell_3, e), (\ell_4, f), (\ell_2, d)\}.
\]

Similarly, the set of features for resource \(b\) is given by the set \(B\).

\[
B = \text{Features}_{LD}(b) = \{(\ell_4, e), (\ell_4, f), (\ell_5, g), (\ell_2, d)\}.
\]

In our experiments, we manually remove features with functional properties, properties that can have only one (unique) value for each resource, i.e., there are no two distinct resources with the same property value (e.g., identifiers).

*Similarity Metrics.* In the present experiment we use the 3w-Jaccard binary similarity measure\(^{16}\) between two RDF nodes based on their corresponding feature

---

\(^c\)https://jena.apache.org/
\(^d\)https://www.w3.org/TR/sparql11-query/
sets. Intuitively, the more features two nodes have in common, the more similar they are. Let $x$ be the size of the set of features shared by $a$ and $b$, $y$ the size of the set of features only present in $A$, and $z$ the size of the set of features only present in $B$. The 3w-Jaccard similarity is defined to weight higher common features, and weight lower discriminating features, i.e., those only present in $A$ or $B$.

$$S_{3W-JACCARD}(a, b) = \frac{3x}{3x + y + z}, \text{ with } 0 \leq S_{3W-JACCARD}(a, b) \leq 1.$$ 

Example 1 Consider two drugs, namely, Phenacetin (identified by dbRank:DB03783) and Acetaminophen (identified by dbRank:DB00316), with a subset of their relations including: label, brand name, dosage form, and target proteins. As Figure 3 shows, these two drugs share some relations with the resources "Humans and other mammals"@en, dbBank:capsuleOral, dbBank:liquidOral, dbBank:analgesics, Non-narcotic, and dbBank:target-20.

Based on the relations shown in the graph, the feature vectors for dbRank:DB03783 and dbRank:DB00316 can be generated. Thus, the similarity between these two drugs is given by the 3w-Jaccard similarity of the feature vectors.

$$S_{3W-JACCARD}(\text{dbRank:DB03783}, \text{dbRank:DB00316}) = \frac{3 \times 5}{3 \times 5 + 6 + 5} = 0.5769$$

*We have experimented with other measures such as information content and Jaccard, but 3w-Jaccard performed best with the presently used data. However, combinations of different similarity measures may lead to further improvements and lower sensitivity to similarity thresholding, therefore related studies are an important aspect of our future work.*
**Prediction of Side Effects**

All feature vectors and similarities between every pair of drugs are computed and stored in an in-memory drug similarity database (see Figure 1) optimized to support on-line querying. After this step, the off-line phase of the system is completed. Conceptually, this database represents a vector space model where each drug is represented as a dot in a \( n \)-dimensional space (where \( n \) is the maximum cardinality of the feature vectors). Given a drug \( x_i \), the neighborhood of \( x_i \) is represented by the closest drugs (i.e., the ones with higher similarity). We use \( k \)-NN algorithm\(^{17}\) to extract such neighborhoods. In the presented experiments we consider neighborhoods of size \( k = 50 \), and further filter the neighbors according to different thresholds over the similarity between \([0.0 - 1.0]\) with incremental steps of \(0.1\).

Side effects are propagated from one drug to its closer neighbors. That is, for a given drug \( x_i \), the system collects the already known side effects of the drugs in the neighborhood of \( x_i \). The side effects are combined and ranked according to their co-occurrence (frequency) to determine the link’s weight. Let \( W_{UL} \) be the vector with the distance of \( x_i \) to each neighbor, \( L_{UU} \) be the sum of the distances from \( x_i \) to all its neighbors, \( \sum_{w_i \in W_{UL}} w_i \); and \( f_L \) the vector of the relative frequencies for a given side effect \( s \) in all the neighbor drugs. To compute the weight that the side effect \( s \) will have over \( x_i \), we use the average mean formula:

\[
s_{weight}(x_i) = \frac{1}{L_{UU}} W_{UL} f_L.
\]

During our evaluation, for each drug in our knowledge graph, we produce as many predictions as actual side effects the drug has in SIDER, and check whether we can retrieve the same set of side effects for a given drug. Therefore, to predict the set of side effects for a given drug, we change the vector \( f_L \) accordingly to each side effect in SIDER.

**Results and Discussion**

In this section we first describe the data and the methodology we used for evaluating our results. The results achieved are summarized then, including comparison to related works. We provide examples of selected results, and eventually discuss the benefits and drawbacks of our work in relation to the state of the art.

**Evaluation Data Set and Methodology**

The comparative evaluation of our approach to adverse drug reaction discovery was based on the SIDER data set as available in the Bio2RDF project. Basic statistics about the knowledge graph used in our work, is provided in Table 2.

<table>
<thead>
<tr>
<th>Table 2. Basic statistics about the SIDER dataset used.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of drugs</td>
</tr>
<tr>
<td>Number of side effects (i.e.,) ADR</td>
</tr>
<tr>
<td>Number of drug-side effect relations</td>
</tr>
<tr>
<td>min / max / avg number of side effects per drug</td>
</tr>
<tr>
<td>min / max / avg number of drugs per side effect</td>
</tr>
</tbody>
</table>

We used the SIDER drug-adverse effect relationship instances as a gold standard. We performed leave-one-out cross-validation of our approach, measuring various scores to assess the predictive power of our approach by training it on a subset of the gold standard drugs and testing it on the remainder in an iterative manner. Such an evaluation method is commonly used for assessing the performance of ADR discovery systems\(^{10-14}\) and thus provides convenient means for direct comparison with related state of the art.

For the performance evaluation we use specific evaluation metrics for multi-label learning, which are different from the ones used in traditional supervised learning\(^{18}\). Let \( p \) be the size of the set of drugs, thus, for each drug \( x_i \) with \( 1 \leq i \leq p \) we have a set \( Y(x_i) \) of actual side effects (from SIDER), and a set \( G(x_i) \) of predicted side effects using our method. We compute the following four measures for evaluating the results regardless of the ranking of the predictions:
Figure 4. Plot of the results in relation to the similarity threshold.

\[
\text{accuracy}(A) = \frac{1}{p} \sum_{i=1}^{p} \frac{|Y(x_i) \cap G(x_i)|}{|Y(x_i) \cup G(x_i)|}, \quad \text{precision}(P) = \frac{1}{p} \sum_{i=1}^{p} \frac{|Y(x_i) \cap G(x_i)|}{|G(x_i)|}, \quad \text{recall}(R) = \frac{1}{p} \sum_{i=1}^{p} \frac{|Y(x_i) \cap G(x_i)|}{|Y(x_i)|},
\]

and F1-score, which is known as the harmonic mean of precision and recall. Furthermore, we computed three scores that reflect the prediction ranking and the extent to which the methods produce not only good, but also highly-ranked good results, namely: average precision (AP), as defined in \textsuperscript{19}; TopK score, which is the relative frequency of drugs having at least one known (i.e., gold standard) side-effect which is ranked among the top K high scoring side-effects according to a prediction; P@K score stands for the precision at K, i.e., precision computed only among top K ranking side effects per drug (we used K \in \{3, 5, 10\}). We compute the average values of the particular measures across all drugs with non-empty set of ADR predictions made, which is an approach common to the works we compare ourselves to in the next section.

Among the measures we used, the TopK and P@K are arguably the most accurate scores in terms of evaluating the benefit of ADR discovery for certain types of end users like clinical practitioners. As explained in \textsuperscript{20}, these scores are easily grasped by non-informaticians and are therefore apt for explaining the reliability of the system to them. Moreover, in settings where quick decisions are needed, like in clinical practice, users do not tend to perform comprehensive search among many possible alternatives to find the relevant ones.\textsuperscript{20} The TopK and P@K scores reflect the likelihood that such users will find relevant results very quickly at the top of the list of possibly relevant results. For users who require a comprehensive study of possible ADRs (e.g., pharmaceutical researchers), good evaluation measures across the whole range of predictions are important as well, though.

**Summary of the Results**

Figure 4 shows how the core evaluation metrics depend on the choice of the similarity cut-off threshold.

Although using a non-zero cut-off results in reduced number of drugs for which we are able to predict side effects by propagation, there are significant gains in all measures up until the 0.6 threshold. After that, the slight gain is balanced by a drastic reduction in the number of drugs with predictions, therefore we decided use 0.6 as the preferred cut-off threshold.

Table 3 compares our results achieved for the best similarity threshold (i.e., 0.6) to recent related approaches (i.e., those that used SIDER in a cross-validation experiment). The random baseline assigns |Y(x_i)| ADRs to all drugs x_i in a random manner, where Y(x_i) is the set of actual ADRs for the drug x_i in SIDER. The second line in Table 3 corresponds to our approach. The methods that support ranking of the side effects predicted for particular drugs are prefixed by + in the table.

\textsuperscript{f} Similarly to \textsuperscript{12}, we used K \in \{1, 5\}, but we report a relative instead of absolute frequency as it provides for a better comparison.
The method\(^{11}\) only provides box plots with comparable measures and the value given here is the best mean as observed in the reported plots. The last row of Table 3 presents relative improvement of the presented method achieved over the best previously existing method.

### Table 3. Comparison of the results with related methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>AP</th>
<th>Top1</th>
<th>Top5</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>random baseline</td>
<td>0.0198</td>
<td>0.0195</td>
<td>0.0196</td>
<td>0.057</td>
<td>0.0266</td>
<td>0.103</td>
<td>0.01</td>
</tr>
<tr>
<td>Fujitsu/Insight method</td>
<td>0.5951</td>
<td>0.5419</td>
<td>0.5606</td>
<td>0.6349</td>
<td>0.5702</td>
<td>0.9532</td>
<td>0.4141</td>
</tr>
<tr>
<td>Atlas and Sharan (2011)(^{10})</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.3468</td>
<td>0.6344</td>
<td>N/A</td>
</tr>
<tr>
<td>Pauwels et al. (2011)(^{11})</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>ca. 0.3</td>
</tr>
<tr>
<td>Yamanishi et al. (2012)(^{12})</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.4255</td>
<td>0.7006</td>
<td>N/A</td>
</tr>
<tr>
<td>Zhang et al. (2015)(^{13})</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.5134</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Zhou et al. (2015)(^{14})</td>
<td>0.565</td>
<td>0.24</td>
<td>0.337</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Relative improvement: +5.33\% *+125.79\% +66.35\% +23.67\% +34.01\% +36.05\% +38.03\%

The P@K results are given in a separate Table 4 since no related approach reports these measures. The best results in both tables are highlighted in bold font.

### Table 4. P@K results.

<table>
<thead>
<tr>
<th>Method</th>
<th>P@3</th>
<th>P@5</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>random baseline</td>
<td>0.0179</td>
<td>0.0213</td>
<td>0.0219</td>
</tr>
<tr>
<td>Fujitsu/Insight method</td>
<td>0.6105</td>
<td>0.6239</td>
<td>0.6305</td>
</tr>
</tbody>
</table>

### Examples of Results

Examples of top five results according to the F1 and P@5 measures, respectively, are given in Table 5. There is substantial overlap between the top positions of these lists, therefore we give the top five drugs for the F1 scores, and then the top five drugs out of top 30 of those that performed best at the P@5 score but were not in the top-F1 list. We can see some relatively frequent drug types as best performers, such as barbiturates, anti-histamines or NSAIDs in Table 5. This may indicate that our method can provide very good results for certain drug classes, possibly based on some of their inherent features in the Bio2RDF data. This insight is further supported by analyzing the numbers of drugs that perform well in terms of precision at 3 or 5, as depicted in Figure 5.

### Table 5. Examples of top-scoring drugs.

<table>
<thead>
<tr>
<th>Top-F1</th>
<th>Drug</th>
<th>Drug type</th>
<th>F1</th>
<th>P@5</th>
<th>Top-P@5</th>
<th>Drug</th>
<th>Drug type</th>
<th>F1</th>
<th>P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secobarbital</td>
<td>barbiturate</td>
<td>0.9825</td>
<td>1.0</td>
<td>Etodolac</td>
<td>NSAID</td>
<td>0.7059</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carbinoxamine</td>
<td>antihistamine</td>
<td>0.9764</td>
<td>1.0</td>
<td>Ganciclovir</td>
<td>antiviral drug</td>
<td>0.6916</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diphenhydramine</td>
<td>H1 histamine antagonist</td>
<td>0.9762</td>
<td>0.71</td>
<td>Sulindac</td>
<td>NSAID</td>
<td>0.6489</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydroflumethiazide</td>
<td>diuretic</td>
<td>0.9697</td>
<td>1.0</td>
<td>Keterolac</td>
<td>NSAID</td>
<td>0.6264</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pentobarbital</td>
<td>barbiturate</td>
<td>0.9643</td>
<td>1.0</td>
<td>Lansoprazole</td>
<td>proton pump</td>
<td>0.6016</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The figure shows that although there is relatively large number of drugs that perform rather poorly, there is also disproportional number of drugs that perform very well.

This preliminary analysis may indicate that for some types of drugs, the results of our method can be substantially better than the average values reported above. Also, there may be strong correlations between good results among the top K results and the whole prediction set. After more precisely detecting the types of drugs that exhibit such behavior, our method can be useful in both of the characteristic use cases we have outlined before (i.e., the clinical practice vs. pharmaceutical research).
Discussion

Our results clearly show the superiority of the presented method. We achieve the best performance in all metrics, improve the previous results by up to 125.79%. The main contribution of our method over related state of the art is the combination of the following factors: (1) The presented method achieves better results in the Top1 and Top5 measures than the related approaches (by as much as 36.05%). It also provides the corresponding precision at K scores that help to determine how many of the top results are actually typically relevant (e.g., more than 3 out of top 5). These combined measures allow for better assessment of what can be expected from the system in cases when users may want to process only first few results and still get many relevant predictions. (2) The method does not depend on patient or report data, even though they can be easily integrated (e.g., via the ClinicalTrial.gov data set in Bio2RDF). (3) The method can automatically digest up to 12 data sets related to drugs, their structure, mechanisms of action, etc. (based on Bio2RDF status as of April 2016). This is different from related state of the art methods that would need to write specific pre-processing pipelines for each of the original data sets, or emulate our approach that is capable of using the Linked Open Data versions of biomedical databases in Bio2RDF. (4) There is only one parameter that has substantial influence on the results -- the similarity threshold. And even using the default parameter value determined by our empirical study is robust, as observed in cross-validation experiments.

The most serious limitation of our approach is the comparatively lower number of drugs for which predictions can be made with the presented method due to the drug similarity threshold. At this stage, the limitation can be partly mitigated by relaxing the similarity threshold. For instance, values around 0.2 result in predictions for all drugs we have features for, while the evaluation scores are still relatively competitive (e.g., F1, Top5 and P@5 scores are 0.3804, 0.8413 and 0.4052, respectively). Focusing on the drug classes for which the presented method performs extremely well (such as NSAIDs) can lead to complete mitigation of this limitation in practical applications. For future versions of the presented technology, we have been addressing this issue by more sophisticated propagation methods involving broader neighborhoods in the similarity graph and combinations of multiple complementary similarity measures. This is expected to lead to high quality predictions for all drugs processed by the system.

Conclusions and Future Work

We presented a method for discovery of drug side effects (or, alternatively, adverse drug reactions) based on propagation of known side effects between similar drugs. The drug similarities were computed using features selected from two data sets---DrugBank and SIDER---represented in a common machine-readable format as parts of the Bio2RDF project. The presented method is very flexible in terms of adding new sets of features from other relevant resources represented in Bio2RDF, which is an advantage over many related state of the art approaches. In addition to that, we were able to achieve better results than related systems in seven standard evaluation metrics. Our system produced promising results especially relevant to use cases in which one needs high ratio of relevant results present among top few side effect predictions (e.g., clinical practice where physicians may need to check for possibly dangerous side effects of a drug but do not have time to filter through many possibly relevant options). The results also motivate some interesting areas of future work that could increase practical relevance of the presented method.

Figure 5. Histograms of drug precisions@K.
The most immediate future work stemming from the research presented here addresses the sensitivity of the results to the drug similarity threshold. We aim to investigate the influence of adding additional features (from other data sets available via Bio2RDF) for the similarity computation, which may result in more densely overlapping feature space and thus higher numbers of drugs that make it even over aggressive thresholds. As more features may lead to more noise, we plan to experiment also with various feature extraction techniques to minimize the risk. We also want to experiment with multiple combined similarity measures and incorporate unrestricted propagation of the side effect labels within the computed similarity networks, using the technique first presented in 21. This will effectively mean that a side effect may be propagated not only to direct neighbors in the drug similarity network, but also to nodes further away (if the similarity links are strong enough). Last but not least, we want to perform an in-depth analysis of the performance of our method across specific types of drugs and test its predictive capabilities on drugs not represented in SIDER, following the approach presented in 12.

Acknowledgements

This work has been supported by the TOMOE project funded by Fujitsu Laboratories Limited and Insight Centre for Data Analytics at National University of Ireland Galway. The technology described in this paper is a subject of a European patent pending, No. 15198304.6-1951.

Supplementary Material

We host all experimental data used in this paper at http://bit.ly/AMIA2016KEDI. The materials contain pre-processed versions of DrugBank and SIDER data sets from Bio2RDF v4.0, including fixing of a few syntax errors and a simple documentation.

References

Content-specific network analysis of peer-to-peer communication in an online community for smoking cessation

Sahiti Myneni, PhD, MSE\(^1\), Nathan K. Cobb, MD\(^2\), Trevor Cohen, MBChB, PhD\(^1\)

\(^1\) The University of Texas School of Biomedical Informatics at Houston, TX, USA, \(^2\)Georgetown University School of Medicine, Washington, DC, United States

Abstract

Analysis of user interactions in online communities could improve our understanding of health-related behaviors and inform the design of technological solutions that support behavior change. However, to achieve this we would need methods that provide granular perspective, yet are scalable. In this paper, we present a methodology for high-throughput semantic and network analysis of large social media datasets, combining semi-automated text categorization with social network analytics. We apply this method to derive content-specific network visualizations of 16,492 user interactions in an online community for smoking cessation. Performance of the categorization system was reasonable (average F-measure of 0.74, with system-rater reliability approaching rater-rater reliability). The resulting semantically specific network analysis of user interactions reveals content- and behavior-specific network topologies. Implications for socio-behavioral health and wellness platforms are also discussed.

Introduction and Background

In recent years, the penetration of online communities into everyday lives has been astonishing. Researchers have studied these communities using multidisciplinary methods to better understand human health behaviors [1,2]. Qualitative studies have been conducted to characterize (a) different types of social support embedded in forum communication [3], and (b) quality of information being disseminated in social platforms [4]. A significant limitation of the application of qualitative methods to peer interactions in online social media platforms is that these analyses were conducted on small samples of communication on account of the labor and time intensive nature of manual coding required. Voluminous data accumulating through increasing use of Health 2.0 technologies, such as online communities, require methods that are scalable. To this end, machine learning methods have been used to identify specific features of communication within online communities [5-7]. Classification of conversational and informational questions on Yahoo! Answers has also been attempted using a combination of human coding, statistical analysis, and machine learning [8]. Methods of distributional semantics have also been combined with machine learning algorithms to classify consumer health webpages of based on language use patterns [9]. However, these studies ignore the structure of communication (who communicates with whom), which is important to understand peer influence on behavior. Quantitative network modeling has been the most widely used technique to describe and visualize behavioral diffusion, communication structure, and peer influences in social communities [10,11]. However, most of these studies do not consider communication content.

In this paper, we describe a semi-automated method that involves (a) capturing the semantic nuances of communication in an online community using a variant of Latent Semantic Analysis (LSA) [12], and (b) using content-specific classification to reveal structural variations underlying the communication patterns of users of a health-related community. LSA is a distributional semantics method that provides us with the capability to derive relatedness measures between terms from unannotated text. This is accomplished by representing the terms in a high dimensional vector space. The coordinates of a term vector in semantic space are determined by the distributional statistics for this term, such that similar vector representations are created for terms that occur in similar contexts [13]. Evidence suggests that the semantic relatedness measures derived using distributional semantics techniques agree with human estimates, and can be used to obtain human-like performance in a number of cognitive tasks [12,14]. Studies have used LSA to automate the coding of communication content among group members to assess team cognition [15,16], suggesting the applicability of the method for communication analysis at scale. Other studies have also established the utility of other distributional representations for analysis of social media postings [17,18]. In this paper, we apply LSA to the model content and structural patterns underlying user communication in QuitNet, an online social network designed to promote smoking cessation. The paper proceeds as follows: (a) firstly, we present an overview of the materials used in the study, (b) secondly, we describe automated analyses conducted on the QuitNet dataset, (c) thirdly, we conduct network analysis to characterize content-specific user communication patterns in QuitNet, and (d) finally, we conclude the paper with implications of our findings for design of digital health platform that leverage the power of social connections.
Materials and Overview of Methods

QuitNet is one of the first online social networks aiming to promote health behavior change, and has been in continuous existence for the past 14 years. It is widely used, with over 100,000 new registrants per year [19,20]. Previous studies of QuitNet indicated that participation in the online community was strongly correlated with abstinence [21]. Communication among QuitNet members can occur through private email, one-to-one messages in public threaded forums, and public chat rooms. The data set studied in this paper was drawn from a previously studied quality improvement database [19], and is comprised of de-identified messages in the public threaded forums, in which participants post messages and reply directly to each other. A database of 16,492 de-identified public messages from between March 1, 2007 and April 30, 2007 was used in our study. All messages are stripped of identifiers but recoded for ego id (the individual posting the message) and alter id (the individual whose message is being replied to), date and position within the thread.

The main pre-requisite for the semi-automated method described in this paper was the development of an annotated dataset that provides qualitative characterization of QuitNet user communication. Grounded theory based qualitative analysis [22] was conducted to identify QuitNet communication themes, and literature review was subsequently conducted to identify the theoretical roots underlying these themes. Further description of this analysis is beyond the scope of the paper, but additional details can be found here [23]. Here we present a summary of pertinent methodological details and results of the qualitative analysis. An initial subset of 795 messages was coded manually until thematic saturation, utilizing grounded theory techniques - open coding, axial coding, and constant comparison. Messages were classified into 12 themes: ‘Social support’, ‘Cravings’, ‘Traditions’, ‘Quit Obscure’, ‘Teachable Moments’, ‘Quit Readiness’, ‘Conflict’, ‘Relapse’, ‘Quit Progress’, Family and Friends’, ‘Virtual Rewards’, and ‘Pharmacotherapy’.

In the current work, distributional semantics and machine learning were used to associate unannotated messages with communication themes from this qualitative analysis. The complete dataset was then further analyzed using network description and modeling packages, to understand theme-specific structural patterns of user interactions. In the following sections, we present the methodological details and subsequent results of the automated text analysis (Step One) and network modeling studies (Step Two) in the context of QuitNet.

Step One: Automated Text Analysis of QuitNet Communication

The methods of automated text analysis we have employed infer measures of the relatedness between passages of text from the distributional statistics of terms in a large text corpus. Based on our previous work [24], we concluded that the distributional information in our QuitNet corpus was insufficient for the automated derivation of meaningful measures of semantic relatedness between terms. Therefore, we drew on distributional information from the Touchstone Applied Science Associated (TASA) corpus [25], a collection of 37,657 articles designed to approximate the average reading of an American college freshman. We used LSA [14] to derive vector representations of terms in the TASA corpus, such that terms with similar distributions would have similar vector representations, with similarity between vectors measured using the cosine metric. This corpus has been widely used in distributional semantics research, and when applied to this corpus LSA has been shown to approximate human performance on a number of cognitive tasks [12]. LSA was performed using the open source Semantic Vectors package [26]. The log-entropy weighting metric was used, and terms occurring on the stopword list distributed with the General Text Parser software package [27] were ignored. This stopword list consists of frequently occurring terms that carry little semantic content.

Subsequently, representations of the messages in the QuitNet corpus were generated by adding the vectors for the terms they contain, and normalizing the resulting message vectors (we will refer to these vectors as TASA-based QuitNet message vectors). Representations for terms in the QuitNet corpus were then generated by adding the message vectors for each message they occurred in, and normalizing the resulting vector. Subsequently, a second set of message vectors was generated, which we term QuitNet message vectors. A pictorial depiction of the vector generation is presented in Figure 1. We utilized this approach in order to ensure that terms present in the QuitNet corpus, but not in the TASA corpus, could obtain meaningful vector representations on account of their having similar distributions to terms in this corpus that did occur in the TASA corpus. This approach is similar in nature to the reflective approach that we have utilized previously to infer associations between terms that do not co-occur directly [28], and provides a convenient means to combine distributional information from two disparate corpora (TASA and QuitNet in our case).
Inspection of the nearest neighbors of key terms from the QuitNet corpus revealed that the measurements of semantic relatedness derived using this approach were intuitive and readily interpretable. In order to use these generated vectors to support automated coding of QuitNet messages, we used a k-nearest neighbors (kNN) approach. For each message, the system provided a ranked list of codes based on pre-assigned manual codes to the nearest neighbors. The score for a particular code was obtained by adding the cosine measures of the nearest neighbors corresponding to that code. The cosine measure represents the relatedness of the message to its nearest neighbor. All of the coded messages other than the message in question were considered (leave one out cross-validation). Figure 2 illustrates the scoring procedure for each theme. As shown in the figure, the five nearest neighbors to message 10515456 were retrieved. For each of the themes, a score was calculated by adding up the cosine values of the nearest neighbors to which the theme was attached. For instance, the score for ‘Quit Readiness’ was obtained by adding the cosine scores of the nearest neighbors (10449020 and 10581825). These scores were used in the next stages to fine-tune the system for accuracy and reliability as explained in the next section of the paper.

Experimental Setup

Using the LSA technique described in the methods section, vectors representing all of the messages in our QuitNet dataset were generated. We then conducted three experiments to evaluate the extent to which these vector representations could be used to accomplish the automated analysis as explained below.

**Experiment 1: Evaluation of the system accuracy**

**Methods:** The 790 manually coded messages were again coded by the automated classification system. The system returns a scored list of codes for each message (see Figure 3). However, many messages are coded with a small number of codes. So some cutoff point is required for meaningful evaluation. In the preliminary experiment we used ranking as a cutoff, but subsequently based the cutoff on association strength. In the ranking-based code assignment, we had the system rank the codes at multiple levels (e.g. top 2, top 4) as seen in Figure 3. Threshold-based cutoffs were also tested at various levels of the association strength. For instance, at 30% threshold only the codes with a score greater than 30% of the highest score were retained. As shown in Figure 3, when a 30% threshold was applied, codes with score greater than 0.3*highest score (0.3*4.182=1.255, ‘Social support’, ‘Quit Readiness’) were retained. The recall and precision of the system in assigning thematic codes was calculated at various levels of threshold and ranking. Based on error analysis, messages with low-level codes “miscellaneous” and “game” were excluded from the dataset because the message content is not amenable to content-based analysis. The code “game” was assigned to those posts where QuitNet members play a word association game with each other to engage themselves in an activity to curb the cravings. However, single word postings such as this cannot be dealt with by the system effectively as individual message content does not provide sufficient semantic context to interpret the purpose of these words. Similarly, messages that were coded as “miscellaneous” were also excluded because the content does not relate to any of the smoking cessation related themes. The experiment was then repeated with the dataset excluding the messages that belong to the “miscellaneous”, and “game” categories, leaving 533 messages.
Figure 2. Overview of the scoring and optimization procedures used for automated classification system
All: list of all the codes that the system assigned to the message in question, Top 2: ranking cut-off that retains the codes with the top 2 highest scores, Top 4: ranking cut-off that retains the codes with the top 4 highest scores, Threshold of 30%: cut-off based on association strength that retains the codes with scores greater than 30 percent of the highest score, Threshold of 50%: cut-off based on association strength that retains the codes with scores greater than 50 percent of the highest score.

Results
The graph in Figure 3 provides the optimum F-measure estimating system accuracy. The average recall and precision were calculated to be at 0.77 and 0.71 respectively for the themes when considering 5-nearest neighbors at a threshold of 50%. The F-measure was found to be 0.74 in this case.

Figure 3. Automated classification system accuracy (F-measure) at a Threshold of 50%
Top X (X=2, 3, 4) indicates the ranking-based cutoff, Th(Y), Y=40%-80% indicates association strength based threshold, NN (Z) (Z=1, 3, 5, 10, 20) indicates the number of nearest neighbors retrieved by the system to formulate scores for code assignment.
Experiment 2: Evaluation of the system reliability

Methods: A separate dataset of 100 messages was coded by two researchers using the themes that emerged from the grounded theory analysis. The same data set was fed to the automated classification system that was optimized for better accuracy in the previous experiment. The system assigned themes to these 100 messages which were then used to calculate human-system reliability using Cohen’s Kappa measure.

Results: The inter-rater reliability was calculated be 0.83, and the system-rater reliability was averaged at 0.77. Surprisingly, results indicated that the system agreed better with second coder, who coded just these 100 messages, than with the coder of the initial 790 messages upon which the system was trained.

Conclusion: The reliability measures obtained in this experiment indicate that the average agreement of the system with human raters for QuitNet themes approached the agreement between human coders.

Experiment 3: Incorporation of outside semantic information

Methods: Initially, LSA was performed on QuitNet corpus directly to generate message vector representations without TASA pre-training. Then, the vector generation process outlined in Figure 1 was utilized to generate QuitNet message vectors with TASA pre-training. The reflective nature of the method shown in Figure 1 ensures that terms present in the QuitNet corpus, but not in the TASA corpus, would obtain meaningful vector representations. Then, kNN was applied to both the QuitNet vector representations with and without TASA-pre-training separately. F- measures were used to compare the effect of incorporation of TASA corpus to derive QuitNet message vector representations.

Results: Using kNN (k=5) without and with TASA pre-training, the F-measures were calculated to be is 0.53 and 0.74 respectively. Table 1 shows the effects of incorporating TASA on the nearest neighbors of the term “craving”.

Table 1. Most closely related terms to term “craving”.

<table>
<thead>
<tr>
<th>Without TASA</th>
<th>little; him; update; came; ever; stayed; gone; still</th>
</tr>
</thead>
<tbody>
<tr>
<td>With TASA</td>
<td>cigarette; nicotine; crave; craves; smoker; habit; chantix; cig</td>
</tr>
</tbody>
</table>

Conclusion: The accuracy measures obtained in this experiment emphasize the importance of the incorporation of external world knowledge when analyzing social media interactions to negotiate issues with lack of semantic context on account of terse text and community-specific jargon.

Step Two: Large scale theme-specific network analysis of QuitNet communication

Methods: Our entire database of QuitNet messages, consisting of 16,492 messages, was processed by the automated classification system described in previous sections of the paper. The computer-annotated QuitNet data were then used to create theme-specific networks, amenable to analysis using traditional (structural) network analytics to understand theme-specific patterns of social dynamics. The users of QuitNet were classified into five mutually-exclusive classes based on their self-reported abstinence.

Class 1: users who have remained abstinent throughout the study period
Class 2: users who were active smokers throughout the study period
Class 3: users who have relapsed (ex-smoker → active smoker) during the study period
Class 4: users who have successfully quit smoking (active smoker → ex-smoker) during the study period
Class 5: users who have relapsed multiple times during the study period.

Theme-specific network models of the QuitNet data were created by representing users as nodes, and their communication as edges. For each theme-specific network, only edges representing messages annotated with that theme were included. Gephi, an open-source network analysis and visualization software package [29] was used to visualize and analyze these network models. Differences in network structure across themes for multiple user classes were examined using social network metrics [11] that explain node importance within a network (centrality metrics) and network connectedness (network cohesion metrics). Definitions of the metrics used in this study are provided below [11].

1. Degree (or connectivity): The degree of a node is defined as the number of edges incident with the node. If the graph is directed, the degree of the node has two components: the number of outgoing links (referred to as the out-degree of the node), and the number of ingoing links (referred to as the in-degree of the node).
2. Density: The proportion of direct ties in a network relative to the total number possible.
3. Path length: The minimum number of ties required to connect two particular actors, as popularized by Milgram’s famous ‘six degrees of separation’ small-world experiment [30].
4. Cluster: A group of nodes, each of which is connected to at least one other node in the group.
5. Modularity: It is defined as the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random, and therefore can be considered as a measure of the cohesiveness of communities within the network.
**Results:** Inclusive of all themes, QuitNet user communication network of Class 1 is the largest with network diameter of 10, comprising of 1204 nodes and 6093 edges. It is the most connected network, with average degree of 10.121. Class 2 is the second largest network of QuitNet users, with network diameter of 9, comprising of 732 nodes and 1696 edges. It is also the second-most connected network, with average degree of 4.634. However, the Class 2 network has the lowest density (0.006) of all, meaning that the connections constitute a tiny fraction of all potential connections that could be formed in this network. For illustration purposes, Figure 4 presents four content-specific network topologies for Class 1 (abstinent) QuitNet users (same color nodes implies clustering into a sub community, orange indicates a low modularity sub community, purple indicates a high modularity sub community, and node size indicates degree). As can be seen in the figure, network formations based on ‘Social support’ and ‘Quit Progress’ are populous and have higher average degree (7.42, 5.3) compared to ‘Quit Benefits’ and ‘Cravings’ (2.86, 3.42) respectively. This indicates that this group of ex-smokers mostly focused their communication with peers on content related to exchange of social support and monitoring their quit progress, reflecting on how far each of them have come in their efforts to quit smoking. However, modularity of networks specific to ‘Social support’ (0.54) and ‘Quit Progress’ (0.52) was lower compared to ‘Quit Benefits’ (0.67) and ‘Cravings’ (0.65), which indicated the sub communities of users exchanging information about benefits and cravings were strongly connected among themselves in comparison with the rest of the network. On the other hand, the average path length in ‘Social support’ (3.4) and ‘Quit Progress’ (3.6) networks is lower when compared to average path length in ‘Quit Benefits’ (4.4) and ‘Cravings’-related networks (4.1), which implies faster dissemination of the former content types.

![Figure 4](image-url). Content-specific network representations of communication involving Class 1 QuitNet users.
Similarly, Figure 5 presents the content-specific network formations of Class 2 users (active smokers). Unlike the Class 1 ‘Social support’ network, where the high-degree users cluster within a single sub-community, the majority of Class 2 high-degree users within the social support network formed a separate modularity cluster to the rest of the QuitNet users. Further, the ‘purple’ nodes revealed that the ties among high-degree users in Class 2 ‘Social support’ were stronger amongst themselves when compared with their ties with the rest of the nodes in the network. The average path length across all content types was approximately equal at 4.05, except for in the ‘Cravings’ network which at 5.3 is higher than this path length for Class 1 users indicating lower efficiency in information dissemination. Among all content types indicated in Figure 5, the ‘Quit Benefits’ network of active smokers had highest number of isolates and least number of nodes, making it the sparsest and least effective network.
Figure 6 presents network formations involving QuitNet users who have relapsed (Class 3) and were successful quitters (Class 4) during the study period. Interactions involving Class 3 users resulted in a network with diameter of 6 comprising of 116 nodes and 111 edges. It is the least connected network with average degree of 1.914. In contrast, Class 4 user interactions resulted in a larger network with 456 nodes and 763 edges, with an average degree of 3.346. Content-specific topological analysis of user interactions related to ‘Traditions’ indicated that the Class 4 users formed a single cohesive network with higher modularity (0.68), when compared to Class 3 users exchanging similar content (0.17). This implies that the emergent traditions within QuitNet may play an important role in successful quitting. The same trend of higher modularity among Class 4 QuitNet users was found in the network describing their interactions related to ‘Quit Readiness’ (0.79), however this network had high degree users form the highest modularity sub community (purple color) as compared to the rest of the nodes (orange color).

**Figure 6.** Content-specific network representations of communication involving Class 3 and Class 4 QuitNet users

**Limitations and next steps**

The automated methods and subsequent network analytics described in this paper were based on qualitative analysis of a small data sample of 795 messages that were coded until thematic saturation. However, it may be the case that the remainder of the dataset contains additional themes that were not captured. The rapid growth of digital technologies will further complicate this issue, as it will generate a data deluge of millions of messages transmitted over the Web and mobile media. The QuitNet dataset considered in our analysis was recorded in 2007. For future studies, we will obtain further data drawn from recent and larger datasets. However, we believe that the findings from the reported data on human behavior are still relevant, as the basic tenet of forum-based communication (structure and logistics) remains the same. In addition, use of recently emerged methods of distributional semantics (e.g. Reflective Random Indexing [28], neural word embeddings [31]) will be considered in our future studies. The network analysis conducted in this paper was limited to description and visualization of user interactions using five metrics. A more extensive approach toward network analysis could be adopted to identify network motifs [32], integrate social influence models [33,34], and model topological evolution over time [35].
Conclusions and Discussions

This paper describes two studies which focus on the analysis of online social network communication using automated text analysis methods from distributional semantics and network analysis for describing content-specific topology differences. The key contributions of this paper are as follows - 1) it introduces and validates a method to extend qualitative analysis to large datasets; 2) it provides a proof-of-concept for full-scale content-specific network analysis of user interactions in an online health-related community; and 3) it demonstrates the use of automated text analysis methods as a bridge between the qualitative and network components of online social media analysis.

Given the voluminous nature of online social network data, it is important to develop automated methods that can address the large quantities of free text data that are available online. While the performance of the automated method is reasonable, the accuracy of the system may be further increased by adopting more sophisticated machine learning algorithms. Importantly, the automated method provides a tractable and intuitive mechanism that facilitates code assignment to all the messages in the dataset, thus enabling qualitative analysis of large datasets. The reflective approach adopted in the automated classification system scales in linear proportion with the size of the dataset and therefore the automated system can be applied to analyze millions of messages exchanged on social media platforms. In addition, incorporation of external semantic information enhances the applicability of distributional models to social media text analytics. A significant implication of large-scale qualitative analysis of online social media interactions is that it facilitates the inclusion of semantic content into network models of online communities. Content-specific network analysis of QuitNet revealed attributes that describe nodal importance and network cohesion of communication themes across multiple behavioral states related to smoking cessation. In turn, such understanding will allow us to develop user-information interactions that facilitate efficient information dissemination, robust network formation, and targeted support within technology platforms such as QuitNet [24,34,36].

In summary, methods that facilitate automated extension of granular qualitative analyses, and population-level visualization of the results can extend the research and application frontiers of social media, thereby further enhancing their positive impact on health-related behaviors.

Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number 1R21LM012271-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are indebted to Tom Landauer for providing us with the TASA corpus, and contributors to the Semantic Vectors open source package, in particular Adrian Kuhn and David Erni, the contributors of the sparse SVD implementation we used for LSA.

References

DisVis: Visualizing Discussion Threads in Online Health Communities

Drashko Nakikj, MS, MA, Lena Mamykina, PhD
Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

An increasing number of individuals turn to online health communities (OHC) for information, advice and support about their health condition or disease. As a result of users’ active participation, these forums store overwhelming volumes of information, which can make access to this information challenging and frustrating. To help overcome this problem we designed a discussion visualization tool DisVis. DisVis includes features for overviewing, browsing and finding particular information in a discussion. In a between subjects study, we tested the impact of DisVis on individuals’ ability to provide an overview of a discussion, find topics of interest and summarize opinions. The study showed that after using the tool, the accuracy of participants’ answers increased by 68% (p-value = 0.023) while at the same time exhibiting trends for reducing the time to answer by 38% with no statistical significance (p-value = 0.082). Qualitative interviews showed general enthusiasm regarding tools for improving browsing and searching for information within discussion forums, suggested different usage scenarios, highlighted opportunities for improving the design of DisVis, and outlined new directions for visualizing user-generated content within OHCs.

Introduction

In recent years, online health communities (OHC) have emerged as an important complement to the traditional health care. According to the 2009 Pew Report, close to a 1/3 of all American adults had accessed social media for questions related to health. More than 1/2 of e-patients (patients relying on online medical services) living with chronic disease consume user-generated health information and 1/5 participate in generating that type of information. In response, there emerged a large number of OHCs such as breastcancer.org, TuDiabetes.org, PatientsLikeMe.com, to name a few. Studies showed that these forums serve many purposes, from sharing information and increasing knowledge, to emotional and social support, to connecting with patients with similar health challenges, among others. However, most of these forums continue to rely on text-based discussion boards as their main platform for achieving these goals, thus presenting rich repositories of text-based information.

Because of the amount of information collected within OHCs, satisfying users’ information needs can be a complicated and time consuming process. For example, based on the data collected in 2014, an active forum for diabetes self-management, TuDiabetes.org, has over 36,000 members and includes almost 23,000 discussions on broad range of topics. Many of these discussions include dozens of members and hundreds of posts, and last for several months or years, transforming the initial topic of the discussion in multiple directions. Therefore, finding relevant information and forming an impression of the topics discussed and opinions expressed within a single discussion thread can be challenging. While in many domains this can create inefficiencies and lead to frustration, in the context of health information it can lead to particularly disastrous effects, such as participants arriving at wrong conclusions to the detriment to their health.

To address these issues, the long-term objective of this research is to develop novel discussion visualization tools that will help members of OHCs to navigate discussion threads, quickly form an impression of what is being discussed, and identify the most interesting and relevant parts of the conversation. In this paper we present the prototype of the DisVis discussion visualization tool and discuss the results of its evaluation study with individuals recruited from TuDiabetes. In this between subjects study, the participants in the intervention and the control group were asked to review a single discussion thread and were asked to generate an overview of discussion topics and participants, find topics of interest, and summarize opinions while using DisVis (experimental group) or the original discussion board representation (control group). We compared individuals’ answers to a gold standard, generated by either members of the research team or a healthcare professional; the accuracy of individuals’ answers and time to answer were captured and analyzed. The evaluation session concluded with short qualitative interviews assessing individuals’ subjective impressions from the tool and opportunities for improvement. We will present the detailed results from this study and their implications for the design of new platforms for OHCs in the remainder of the paper.

Related Work

OHCs have become a topic of active investigation within several research communities, with the particular attention on information seeking practices of their members. Previous research within OHCs found that users expressed...
higher levels of satisfaction with their experience when their information needs were matched with appropriate information. Along these lines, previous studies explored common characteristics of health information seeking. These studies suggested that individuals prefer answers by others matched on profile similarity, highlighted the importance of context when posing queries, and suggested that search results should be personalized based on user's medical history. Consequently, recent trends in the research on OHCs increasingly favor new computational solutions for optimizing access to information, expertise and interest matching, mining members' contributions over time to infer their credibility, and using automated ways to extract personal information from members' contributions. In regards to information seeking in OHCs, Nambisan suggested the need to focus on developing tools that make information seeking more effective and efficient.

Beyond health-oriented discussion forums, other domains have explored the question of accessing discussion information through visualization in great depth, including visualization of enterprise flash forums, e-mails, and newsgroups. For example, ConVis is a system that enables the user to quickly form a topic-post-sentiment-author chain connection when inspecting the discussion. ForumReader is a tool developed for exploration and analysis of flash forums based on the thumbnail representation of the discussion. Themail builds a visual display of e-mail based interactions over time, while Newsgroup Crowds graphically represents the population of authors in a particular newsgroup in Usenet and AuthorLines visualizes a particular author's posting activity in the same platform. Other researchers have looked at preserving and visualizing the tree structure in the discussions using the sunburst radial space filling approach, icicle tree layout and adding reply-based sociogram view. Other approaches to visualizing large amounts of information relevant for this research were explored within topic based visual text analysis, large network data visualization with rich user interaction and visualizing social roles. Finally, previous research explored ways to visualize opinions about a product and opinions about ideas.

Despite this ongoing research on visualizing user-generated content collections, addressing information overload in OHC forums with visualization solutions has received limited attention in the past. Within this domain, previous research focused on topic modeling and network visualization to explore patients' experiences and on modeling interpersonal interactions and medication use, as well as health behavior and symptoms polarity. Another direction explored by previous research focused specifically on supporting community moderators. For example, VisOHC is a tool for OHC moderators that integrates visualizing the forum, aggregating diverse dimensions of conversational threads.

Although these systems and tools represent a rich body of knowledge about forum/discussion visualization, few of them focused specifically on helping members of OHCs overview discussion threads within their community, find topics of interest, and synthesize different opinions on these topics. As a result, the goal of this work was to design a simple discussion visualization tool that was tailored to discussions within OHCs, and to evaluate it with members of one such community. In the spirit of technology probes, we intended this prototype as a tool for proving a concept: testing the success of basic design approaches and eliciting user feedback.

**Methods**

**The Prototype: DisVis**

The design of DisVis was informed by our own previous studies of collective sensemaking within TuDiabetes and by review of related literature. Our particular focus was on assisting users in three main directions: 1) forming an overview of the entire discussion thread, 2) finding prevalent topics within a discussion, and 3) summarizing different opinions in regards to the prevalent topics. DisVis interface is logically divided into two main segments: Exploration and Navigation area (Figure 1-B) and Full Content area with a summary (Figure 1-A). These two segments were introduced to allow users to move seamlessly between exploring different aspects of the discussion and reading the text of interest in more detail, respectively. We describe these features in detail below.

**Discussion overview.** The first set of features described here was included to help individuals form a quick impression of the entire discussion thread. The Summary section (2) provides a brief summary of the discussion, including its key descriptors: number of posts, number of participants, duration, and keywords. The Index Panel (4) lists the most prevalent keywords (identified using frequency of keywords within the discussion) and most active contributors (number of posts) in descending order from left to right. The Content-contributor Overview Bar (12) provides the overview of where in the discussion a user talked about a particular keyword. The “k” label provides insight in the distribution of the presence of a given keyword in the discussion, selected from the Keywords in the Index Panel. Analogously, the “u” label does the same for a given user, selected from the Users in the Index Panel. The Timeline (13) gives the overview of when the posts were posted, thus helping to determine the distribution of
activity in the discussion over time. The Posts Length section (9) shows the amount of elaboration in terms of
numbers of words in each of the posts, the line length being proportional to it. The Replies Level section (10) gives
the overview of the users’ depth of engagement – the length of the blue bar represents the depth of the reply.

**Finding topics of interest:** the key feature for enabling this capability is the Keywords in the Index Panel (4). By
selecting a keyword from Keywords, only posts that contain that keyword are displayed in the Discussion section
(3), keeping their sequential order. A second level of filtering topics of interest is introduced in the Posts Keywords
(identified by tf-idf metric where a post is considered to be the document and the discussion the corpus) section (8),
where the user can see which are most important keywords for that particular post.

**Summary of opinions:** the key feature in this category is the Index Panel (4) again. By selecting a keyword from
Keywords and a discussion participant from Participants, the user can filter the posts that mention that keyword and
come from that participant.

![Figure 1](image_url)

A snapshot of the DisVis user interface in use with functionality markers: (1) Category and Heading; (2)
Summary; (3) Discussion; (4) Index Panel; (5) Interaction; (6) Context; (7) Authors; (8) Posts Keywords; (9) Posts
Length; (10) Replies Level; (11) Slider; (12) Content-contributor Overview Bar; (13) Timeline; (14) Time and Date.

In addition to these three main groups of features, the remaining features of DisVis allow for integration between
the two main segments (A and B). The most important of these is the Slider (11). Its cursor can be moved to point to a
given post in its graphical presentation. Here, each post is represented by a vertical line segment, preserving the
sequential order in the standard discussion representation (3); the different characteristics of these segments indicate
their length (line height - 9), keywords (8), author (7), chronological position (13), time and date (14), and depth of
engagement (rectangle height -10). To provide the social and informational context of the selected post, the
Interaction feature (5) shows the set of people the author of that post had interacted the most in the discussion and
through Context (6) it shows the posts that replied to it and the post it replied to.

**Usage scenario.** Elizabeth was recently diagnosed with Type 1 diabetes and was recommended an insulin pump as a
treatment. She has heard about TuDiabetes and decides to visit the forum to learn more about it. She finds a
discussion thread (displayed in the traditional way) that appears to be focused on insulin pumps, but finds that it has
5 pages of posts (a minimum of 41 posts). Elizabeth is not sure whether this is the right thread to focus on and
decides to use DisVis to learn more about who participated and what was discussed. From the Summary section (2),
Elizabeth can immediately see that 19 members of TuDiabetes participated in the discussion contributing 46 posts
over only 2 weeks, and that the main topics included insulin pump, infusion sets, insulin, MDI, basal rates and
She notices that one of the keywords included in the Index Panel (4) refers to the topic she is interested in, “pumps”, and decides to investigate further. Elizabeth also notices that many of the posts have multiple replies by looking at the Replies Level section (10), which suggests to her that participants engaged in a discussion over the topics of mutual interest. This high-level overview gives Elizabeth a good sense of the complexity, richness and duration of the discussion, as well as participants’ engagement. The initial post of the discussion is an appeal from Jennifer for troubleshooting unsuccessful insulin pump treatment and finding alternatives. After reading that, Elizabeth wants to explore Jennifer’s opinions about pumps and whether they changed during the discussion. In the Index Panel (4), she selects “Jenifer” as the user, and selects the keyword “pump” from the Index Panel (4) as her topic of interest. In the Discussion section (3), she sees filtered posts that contain “pump” as a keyword and Jennifer as the author. She then reads those posts and formulates the opinion that Jennifer had on pumps. She does that for 2hoobit1 and other participants. However, she also notices that participants are mentioning MDI as an alternative treatment. Elizabeth goes through an analogous process to find out what they think about MDIs. As a result of these explorations, Elizabeth forms a pretty good idea of the different perspectives on pumps that were expressed in this discussion, and how members’ perspectives changed over time. She also identifies an alternative treatment – MDI, that she can read more about in the future. She makes a mental note of the names of active participants and decides to pay attention to their posts in the future. And she is able to accomplish all of that in a matter of minutes, without reading every single post in this discussion thread.

**Study Design**

**Sample and setting.** The participants of the study were recruited among members of the TuDiabetes OHC using announcements on its home page and with the help of the community leadership. Being a member of this OHC was the only eligibility criterion for the study. The study was conducted on-line, using the Join.me shared screen platform for interacting with a discussion using the tool or within its original environment in TuDiabetes.

**Assessment measures.** There were three types (subsets) of questions used to evaluate the tool: S1, S2 and S3, with a total of 15 questions: a) general overview of the discussion (S1), related to providing a summary of the discussion - finding the key concepts discussed and most active participants, activity of the discussion, etc. (e.g. What were some key concepts discussed here?); b) finding topics of interest (S2), related to the ability to filter posts about a given topic (e.g. Who contributed the most posts that talked about infusion sets?); c) opinion synthesis (S3), related to finding what different opinions are on a given topic (e.g. How would you summarize the general opinion about MDI, do you think most participants think it’s a good idea, bad idea, or is there a split in opinions?). To assess participants’ accuracy, we developed a gold standard answers (GS) for each of the 15 questions. For purely objective questions (e.g.: How many individuals participated in the discussion?) the answers were generated by the members of the research team. For questions that required domain knowledge and synthesis of opinions (e.g.: What were the key concepts discussed in this post?), the gold standard answers were generated by a clinician (a doctoral student in nursing with experience in diabetes self-management). We took the following approach to measure the accuracy based on the developed gold standard. For answers that had a single discrete value, the score could be either 0 or 1 (accurate=1, inaccurate=0) and for those that had several discrete values, we calculated the Jaccard* similarity coefficient between the values provided in the answer and the GS answer. For those that were qualitative and more descriptive, we assessed the similarity by Jaccard to the GS based on the key points covered in the answer compared to the GS (each key point being a discrete value, hence the user’s answer and the GS answer being two sets of discrete values). To assess participants’ time to answer questions, the researchers reviewed audio records of the sessions and measured time (in seconds) from the end of the question to the end of the final answer provided by the participant. The participants were given a maximum of 5 min. to start answering a question. If in that time frame an answer was not provided, their accuracy score would get set to 0 and time to answer would get set to 300 sec (this never happened however).

**Study procedures.** Upon enrollment, the participants were randomized into the experimental condition or the control condition (N=5 for each group). The participants in the control group were presented with a discussion thread within its original environment, TuDiabetes, and were asked a set of questions about the information contained in the discussion. After they answered all the questions, they were presented with DisVis, received a short tutorial, and asked for their initial feedback about the tool. The participants in the intervention group were presented with the same discussion thread within the DisVis interface. After receiving the same tutorial for the tool, they were asked the same questions as the control group. Then, they were asked for their feedback on the design of the tool and their

* Jaccard similarity is a metric for measuring the overlap between two sets of discrete elements which can take a value between 0 and 1. A higher value means a bigger overlap.
experience using it. All evaluation sessions were done through the Join.me platform and audio recorded and transcribed verbatim for analysis.

**Analysis.** To find out whether DisVis had an effect, we performed a two-tailed two sample t-test for each of the two outcomes: time and accuracy. Additionally, we wanted to see the tool’s effect for each type of questions separately: S1, S2 and S3. Due to the small number of questions in each of the question categories, we used non-parametric Mann-Whitney U Test (Wilcoxon Rank Sum Test) to determine the statistical significance of our findings. For the ranking, we used the average performance of each participant for the given set of questions. For the qualitative data collected in the study (participants reflections on the tool and their experience with it) we used inductive thematic analysis. The recorded evaluation sessions were transcribed and open coding approach was taken to help us identify and categorize the prevalent themes from the user feedback. The first author (DN) coded all transcripts independently. The emerging categories were then discussed during meetings between the authors and iteratively refined in a collaborative process.

**Results**

Ten participants were recruited to participate in the study (N=5 in each group), 6 male and 4 female. More than half of the participants were above 60 years of age; diabetes diagnoses included both Type 1 and Type 2 diabetes having experience living with the disease ranging from 3 to 45 years (half of them for more than 20 years). The experience with the TuDiabetes.org ranged from only several months to being present in the community since its inception – almost 8 years. However, the vast majority of the participants were regular visitors of the forum, both for reading and contributing purposes; only one participant self-described as a sporadic user of TuDiabetes.

**Quantitative**

<table>
<thead>
<tr>
<th>category</th>
<th>measurement</th>
<th>group</th>
<th>mean</th>
<th>sd</th>
<th>significance test (α=0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Control</td>
<td>70.16</td>
<td>42.23</td>
<td>(U=4, p=0.095)</td>
</tr>
<tr>
<td>S1</td>
<td>Time (sec)</td>
<td>Intervention</td>
<td>34.73</td>
<td>23.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Control</td>
<td>0.34</td>
<td>0.06</td>
<td>(U=1, p=0.016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>0.59</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>Time (sec)</td>
<td>Control</td>
<td>94.3</td>
<td>41.72</td>
<td>(U=4, p=0.095)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>47.9</td>
<td>43.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Control</td>
<td>0.6</td>
<td>0.22</td>
<td>(U=12.5, p=1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>0.6</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Time (sec)</td>
<td>Control</td>
<td>40.7</td>
<td>16.10</td>
<td>(U=12, p=1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>49.8</td>
<td>36.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Control</td>
<td>0.37</td>
<td>0.19</td>
<td>(U=1, 0.016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>0.77</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time (sec)</td>
<td>Control</td>
<td>65.52</td>
<td>46.77</td>
<td>(t=1.80, p=0.082)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>40.51</td>
<td>26.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Control</td>
<td>0.38</td>
<td>0.28</td>
<td>(t=2.41, p=0.023)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>0.64</td>
<td>0.31</td>
<td></td>
</tr>
</tbody>
</table>

Noticeable differences were observed in the advantage the tool brings over the control group. The tool was shown to highly improve the accuracy of and time to answering questions. Also, participants in the intervention group outperformed those in the control group for some, but not all of the question sets tested. From Table 1 (All) we can see that the average time to answer a question for the control group was 65.52 sec (sd=46.77 and median of 63.8 sec) as compared to 40.51 sec (sd=26.38 and median of 37.6 sec) for the intervention group. The average accuracy score for the control group was 0.38 (sd=0.28 with median of 0.35), and 0.64 for the intervention group (sd=0.31 with median of 0.8). The t-test showed that there is an effect for the accuracy of answering when using the tool: the accuracy when using the tool being 68% higher than without the tool (p-value = 0.023). However, the positive trends in reducing the time to answer questions by 38% was shown to be statistically insignificant (p-value = 0.082). The tool was shown to perform better than or at least as good as the control in all cases, except for time to answer questions from S3. However, statistical significance was observed in only two of the cases in respect to question types: improving the accuracy of questions of type 1 (S1) and type 3 (S3), both with p-value = 0.016.

**Qualitative**

The focus of the qualitative interviews was on the participants’ subjective impressions of the tool and their experience using it. The analysis of the transcripts suggested that the participants’ experiences with the tool, and
with the forum at large, differed depending on their overall goal. In some cases, the participants came to the forum without any specific questions, but rather to look around for what’s new, which included browsing of recent additions to the discussion threads, and skimming of the new content. However, in other cases, the participants came to the forum with a particular question in mind and their interaction focused on identifying relevant information through reading relevant posts and threads, and comparing different opinions on the topic. Notably, these two scenarios were interchangeably utilized across most of the participants. Also, and not surprisingly, these two usage scenarios were associated with very different expectations for how to optimize individuals’ experiences and the desired set of functions within the forum. Below, we describe the participants overall assessment of the DisVis tool in terms of usefulness and usability, as well as the more specific comments from the perspective of browsing and searching for specific information.

**Overall assessment of the DisVis tool.** Overall, the majority of participants agreed with the general premise of the tool, and found its goal of increasing readability of discussion threads and promoting individuals’ ability to explore threads to be promising and potentially useful. P1: “Well, I think it’s a great idea because I’ve had some struggles reading discussions, to be honest...” P8: “Well, for sure the tool is good to have since it gives you a clear idea about what’s going on.” However, together with these positive general perceptions, the participants had many critical comments regarding the tools’ current design. Many thought it was too complicated and overthought and provided too much information, at times, without a clear purpose. P5: “And I back up a little bit more, from my way that I use the site, this stuff would be more noise than signal.” Notably, those who could not find personal use of the tool proposed that it might be most useful for those who started the thread, as well as for moderators, analysts, and researchers. P9: “I mean, if you are a scientist analyzing it [the discussion] and you are trying like to pinpoint things like you’ve been asking me, than I can understand it [the reason for using the tool], but just for an average reader – I don’t think I would use it.”

Finally, the participants were split in their opinions about the tool’s intuitiveness. Most users were able to successfully navigate the tool after the brief tutorial. However, several noted that they would require more experience using the tool to become proficient, which indicated limitations in its intuitiveness. P1: “…there are also I think some struggles at the beginning, but now I think with a bit like getting used to it, it’s not that difficult to use.” P2: “Well, you know I understood the instructions when you gave them, but putting them into practice, I think after I practice with it a couple or three times, I would have a down path, but just based on one run through, I don’t -- I would need more time to learn it and practice with it.” To address this problem specifically, given the number of different features available to users, several participants suggested the use of tool tips to remind them of the purpose of different features.

**Exploratory browsing.** While talking about their experience using the tool, many participants contextualized them within their overall experiences with TuDiabetes and its particular platform. As we mentioned above, these experiences fell within two distinct categories: exploratory browsing and focused information search. We found that many of the specific comments in regards to DisVis’s usefulness were also made in reference to its ability to support either browsing, or focused search.

Most participants made positive comments in regards to the tool’s ability to support browsing. For example, the participants found the overview of the discussion, Summary statistics of the discussion and the Timeline, to be particularly helpful. P2: “I like the dates, I like the date of the -- I really like knowing the timeframe of the thread and the number of participants. All of that summary information is very good.” On the other hand, the Slider did not live up to the expectations. For example, the interactive graphical part of the visualization that involved the Slider, developed for navigational purposes for quick browsing, caused the most negative reactions. This was primarily due to the density of information, small font size used in that area and poor adjustability for different screen sizes, especially for small screens. The participants had split opinions about the list of participants and their number. Some of the participants found those useful for formulating opinion summaries or knowing what positions to expect in the discussion; others found them somewhat irrelevant.

More generally, however, some participants questioned the tool’s attention on a single discussion thread as the only focus in the visualization. When describing their browsing experiences, these participants often talked about the need to see a discussion within the context of the forum, rather than as an isolated instance. P3: “I’m more inclined to look at all the discussions, all the responses.” These participants were likely to perceive individual threads as existing in a symbiotic relationship with other discussions that are on the same topic or have a certain amount of

---

* At the time of the study, TuDiabetes used Ning as its platform; since then the forum switched to Discourse
overlap in the topics they cover, alluding for a broader relevant context when reading a particular discussion. This colored their perceptions of the tool’s focus on a single discussion thread: P7: “...so if you clicked on pump you would see a list here of all of the discussions’ summaries of whatever, of, of whatever is related in, in another discussion, that, that would be good.” These participants wished to see the overlap between different discussions on the same topic, and to compare the topics between discussion threads. P10: “Perhaps instead of, instead of this, maybe, on keywords, if you clicked on insulin or mdi or pump instead of showing where it shows up, here, what other discussions are mentioning pump, insulin and mdi...” Participants were also interested to see forum wide statistics, a more quantitative description of what was going on in the forum: P3: “I might want to compare this to discussions and see what the keywords are for both and see if I could cross reference some keywords. I might also want to know what the community’s trending keywords are.” Finally, the participants proposed the following new features for more efficient skimming of discussions: highlighting keywords of interest in the text, finding the latest posts in the discussion, and labeling the read posts.

Focused information search. According to the participants, the tool was helpful in finding specific information of interest. The most-liked feature for this task was the Index Panel with its Keywords section. P2: “I would love to sort of sort out especially with long threads the ones that really went back to the key points of the initial discussion, like in this case pump and insulin I would probably because I’m an insulin user. I would probably pick on insulin and pump. So it does help me...” However, some participants thought they could not find much use of the Index Panel in real life, although it helped them answer the questions in the study. The participants also proposed specific enhancements to existing features. These included: more prominent highlighting of the selections in the Index Panel, and selection of multiple keywords for the purpose of building more complex Boolean queries.

Further, the participants made many comments regarding desired new features that could improve searching for specific information. These included features for identifying posts of relevance within a particular discussion (as was the case in DisVis) as well as for finding information of interest within the forum at large, or as some participants referred to it “starting from scratch”. P10: “So let’s say I go in, onto the home page, and I have a question about the pump... you know... how do I find it from scratch?! Instead of being right here on this open page that’s already started and everything, how would I get to that page from scratch?? – Do I see five different discussions going on, 20 different discussions and then, what would I do??” For the first case, the participants proposed keywords (terms) that distinguish a discussion from other similar discussions, as well as a keyword cloud. P4: “…I’m in the insulin pump users’ forum so I would already expect there to be discussions about the pump and oftentimes, a lot of the discussions are going to include comments about the infusion set, about the insulin, about the basal rates and multiple daily injections. So I guess the question for me is how could we better define keywords that actually provide more relevance... […] Yeah, I think more about differentiating the discussions [in terms of keywords] because I think a lot of people will come out in these forums and will search for somebody having the same issue or same challenge that they have today.” For the second case, the participants strongly wished for a better search engine, claiming that the current one is simply not delivering the most relevant and recent discussions, and it seems to be more oriented towards individual posts within a discussion, rather than a discussion as a whole. P1: “If you search for a key word it doesn't really show you what you are looking for, like searches for specific answers and I want the overall discussions, so that's not much of help...” In addition, an ability to identify clusters of discussions was highly desired as a feature that can improve the searching process. Finally, the participants liked the idea of exploring different opinions about a topic of interest. However, there was no clear agreement regarding the best way to achieve this. Some wished for a condensed, potentially computer-generated summary; others felt that only reading multiple posts from the same person on a topic of interest could help to synthesize their perspective, an ability which comes through long time presence in the community and reading of the forum.

Discussion

Previously proposed tools for visualizing participant-generated content explored several different perspectives. Many tools focused on providing an overview of the entire volume of the content collected (e-mails or news group postings), rather than focusing on a specific conversation. These included visualizations of interpersonal interactions and topics covered in those interactions; engagement of participants in discussions and individuals’ activity and volume of contributions over time. Other tools took an approach more consistent with the one explored in this work and focused on providing overviews of individual discussions as well as more detailed information about them. Some of these tools placed greater emphasis on finding relevant posts based on topics of interest and participants, while also incorporating information regarding the sentiment of the topic. Others were concerned with providing visual, often scatterplot-based depictions of clustered discussion content. These tools distinguished discussions and threads of conversation within the discussion, highlighted different topics and participants, and
importance of individual posts. The evaluation studies of these tools focused on users’ ability to find topics of interest, summarize, compare and contrast opinions, understand the sentiment about a topic, activity of the participants, their expertise and their interpersonal interactions and relationships. The authors’ activity oriented studies showed that knowing authors behavioral patterns in terms of activity and types of contributions was highly valued for understanding their roles and domain knowledge in the community. On the other hand, the discussion visualization studies found that users particularly liked the ability to find posts that match their topic of interest, which was found useful in answering questions related to summarizing opinions. Here, however, the findings about the importance of authors’ details were dependent on the task study participants were asked to complete.

Consistently with these previous studies, the participants in our study were positive about the general premise of an interactive tool for visualizing salient features of a discussion thread and exploring its different characteristics and attributes. We found that the participants viewed the tool in the context of two typical scenarios of use, corresponding to the two general types of engagement they experienced with the forum, a distinction not considered in the previous research. On one hand, they looked for features that could help to summarize the thread in a bird’s eye view, highlighting such characteristics as most prevalent keywords and most active participants. On the other hand, they particularly liked the possibility to find posts of interest using the Index Panel, which helped in finding specific pieces of information and formulating opinions about a topic. This is consistent with other studies which found that users particularly liked the ability to find posts that match their topic of interest. In addition, the participants valued an ability to view a temporal visualization of the discussion in the interactive timeline, an ability that has not received much attention in prior research, but was considered particularly useful in our study. These positive impressions were also supported by the results of the quantitative assessment: participants in the intervention group who used DisVis interface were able to answer questions related to the discussion thread with higher accuracy and in less time than the baseline. However, together with these positive impressions, the participants made a number of critical comments, particularly in regards to the tools’ intuitiveness and clarity. Most importantly, they found the information display to be too dense and hard to read, and some of the interactive features, for example the Slider, too difficult to operate. This can be related to a finding from another study, where participants underperformed when exposed to both visualization of the discussion and text analytics compared to a standard discussion representation. However, it contradicts our case in which, based on the quantitative findings, the complexity of the tool actually improved performance as compared to the baseline. From the perspective of general browsing, the participants questioned the tool’s focus on a single discussion thread, and desired an overview of topics of interest throughout all discussion threads in the forum. In contrast to our work, this kind of an approach was predominant in VisOHC which focused on OHC moderators. From the perspective of focused search for information, participants found one keyword selection to be limiting and wished for an ability to construct more complex queries. In addition, they wished for an ability to see keywords selected within the Index Panel highlighted in the discussion’s text.

The study helped us to formulate a number of design implications for the next generation of tools for visualizing discussions in OHCs. First of all, the study once again stressed the importance of designing tools that could be used by a variety of individuals regardless of their background, age, specifics of their disease, and computer literacy. Since these tools are likely to be used in an unsupervised fashion and without any possibility for training or instruction, they need to be intuitive and easy to learn. The study also generated a number of specific recommendations for features that could promote the two complementary ways of engaging with information. First, regarding the browsing experience, the desired features included: 1) discussion summary (statistics); 2) larger context, including immediate access to relevant discussions; 3) capabilities to estimate and locate the overlap between the relevant discussions; and 4) keeping track of what was read and what is new. Second, regarding the focused information searching, the desired features included: 1) description of the discussion that distinguishes it from the others; 2) building more complex queries to search a particular discussion and 3) computationally assisted ways to summarize different opinions within the discussion.

In addition, the study generated new design directions for helping individuals engage with user-generated content within OHCs. These include: a) identifying and highlighting relationships between different discussion threads, for example those that include similar topics and b) developing customizable mechanisms to summarize varying opinions on topics of interest. However, in regards to summarizing opinions, the participants lacked clear consensus on whether they were comfortable relying on computer-generated summaries or whether they preferred tools that helped humans to synthesize opinions in a more efficient way. This division is reminiscent of the long-standing argument about the comparative benefits of computational agents and information visualization displays with direct manipulation. Just as this debate came to the conclusion that an ultimate solution requires a combination of both
approaches, we suspect that future tools for supporting online discussion forums should incorporate both computational methods and information visualization techniques.

Study limitations. This study has a number of limitations. First, because the study relied on a real discussion thread captured within TuDiabetes, and because we did not screen participants on their previous exposure to this particular thread, it is possible that some of them saw it before the study. In fact, one of the participants mentioned that the thread looked familiar, but could not recall any specific details. In addition, because the study used only a single discussion thread, it is possible that the findings of the study are specific to this thread and may not generalize to other threads within the same forum or to other forums. In addition, the questions used for assessing the tool’s efficacy may not be relevant to all individuals who frequent OHCs. In fact, few of these questions were found less than relevant by some of the study participants. Additionally, sharing the screen occasionally caused some difficulties with interaction on the participants’ side for both study groups, which was more pronounced for those that needed to interact with the tool.

Conclusion
In this paper, we introduced DisVis – a tool for visualizing discussion threads in an OHC and discussed the results of a study examining DisVis’s impact on individuals’ ability to find information of interest and better understand discussion threads. The study showed that DisVis had a positive overall effect on individuals’ ability to provide overview of the discussion, find topics of interest and summarize opinions. The tool increased the accuracy of the answers by 68% and reduced the time to answer questions by 38% (the latter being statistically not significant). The study also showed that the community members are open to these types of solutions and suggested a set of enhancements and new features for the tool that we plan to address in our future work.

References
Predicting Prolonged Stay in the ICU Attributable to Bleeding in Patients Offered Plasma Transfusion

Che Ngufor, Ph.D., Dennis Murphree, Ph.D., Sudhi Upadhyaya, Nageswar Madde, Jyotishman Pathak, Ph.D. 1, Rickey Carter, Ph.D., and Daryl Kor, M.D.

Mayo Clinic, Rochester, MN, 1 Weill Cornell Medical College, NY, NY

Abstract

In blood transfusion studies, plasma transfusion (PPT) and bleeding are known to be associated with risk of prolonged ICU length of stay (ICU-LOS). However, as patients can show significant heterogeneity in response to a treatment, there might exist subgroups with differential effects. The existence and characteristics of these subpopulations in blood transfusion has not been well-studied. Further, the impact of bleeding in patients offered PPT on prolonged ICU-LOS is not known. This study presents a causal and predictive framework to examine these problems. The two-step approach first estimates the effect of bleeding in PPT patients on prolonged ICU-LOS and then estimates risks of bleeding and prolonged ICU-LOS. The framework integrates a classification model for risks prediction and a regression model to predict actual LOS. Results showed that the effect of bleeding in PPT patients significantly increases risk of prolonged ICU-LOS (55%, p=0.00) while no bleeding significantly reduces ICU-LOS (4%, p=0.046).

Keywords: Blood transfusion, perioperative, bleeding, machine learning, classification.

Introduction

Great progress has been made in improving patient outcomes associated with major surgical operations in recent years. This has been manifested by the reduction in intraoperative and postoperative mortality and the overall reduction in transfusion requirements 1. Despite these advances, bleeding is still the most frequent serious complication during or early after major surgical procedures. For example, bleeding in the immediate postoperative period occurs in approximately 20% of patients undergoing liver transplant and about 12% in patients exposed to cardiopulmonary bypass 2. Bleeding in the perioperative period has been found in many studies to be significantly associated with increased health care resource utilization as reflected by increased intensive care unit (ICU) length of stay (LOS) 3;4, morbidity and mortality 5.

After major surgical operations patients are typically managed in the ICU, thus the duration of stay is an important indicator of the quality of care a patient receives as well as resource utilization. Patients with prolonged ICU-LOS generally comprise a very small proportion of all ICU patients, yet studies have repeatedly shown that they consume a significant share of ICU resources 6;7. Early identification of patients who might stay longer in the ICU can help improve resource utilization and efficiency of ICU care.

Many studies have reported risk factors for prolonged ICU-LOS and bleeding and a few have reported accurate models for early identification of patients who might experience these outcomes 6–8. The majority of these studies have considered the outcomes independently. Little attention has been given to the identification of patients groups with prolonged ICU-LOS or increased risk of bleeding attributable to other important clinical outcomes or interventions. A number of studies have investigated the attributable mortality and ICU-LOS of major bleeding using matched cohort analysis or Cox regression models 3,4,7. In the analysis, variables observed in the intraoperative and early postoperative periods are typically incorporated in the models for improved performance. For example, in Kramer and Zimmerman 7 predictive models for prolonged ICU-LOS that incorporate patient data observed as late as day 5 of ICU admission produced more accurate results than day 1 data. However, as such studies assume the availability of intraoperative and/or post-operative data they are inapplicable when only preoperative data is available. The authors of this study are unaware of any study that has investigated the causal relationship between prolonged ICU-LOS in patients intervened on with a treatment as plasma transfusion, and methods to early identify patients who might experienced these outcomes using only predictor variables observed at the preoperative period.

1Unless otherwise mentioned, any reference to bleeding will be understood as perioperative bleeding.
A key step to identify patients who might stay long in the ICU attributable to bleeding is to assess preoperatively, patients who might bleed during or early after surgery. Across many centers in the US, substantial emphasis is often placed on preoperative screening tests, such as the international normalized ratio (INR): a major driver of decisions about preoperative plasma transfusion. Fresh frozen plasma (FFP) infusions are commonly used to improve coagulation or clotting and are the main therapy option for patients with elevated INR. A large proportion of plasma components are transfused in the perioperative environment, however, they are frequently administered prophylactically in the absence of significant active bleeding. This practice persists despite a growing body of literature questioning its efficacy. Through the use of machine learning methods, a recent study conducted by the authors confirmed previous findings that population-wise, preoperative plasma transfusion (PPT) significantly increases the risk of bleeding, ICU-LOS, re-operation due to bleeding, and other important outcomes for patients undergoing non-surgical procedures.

The goal of this study is to present a causal and predictive framework where in a first step, the causal effect (attributable risk) of bleeding in patients offered PPT on prolonged ICU-LOS is estimated. Then in a second step, a predictive model estimates the risk of (1) bleeding and prolonged ICU-LOS attributable to bleeding in patients offered PPT, and (2) all causes of prolonged ICU-LOS. The predictive risk modeling step is designed as a two level structure where in the first level, a classification model estimates the probability of bleeding given that a patient was offered PPT. In the second level, a regression model is constructed specifically for patients identified by the first level model to be at high risk of bleeding to predict the actual ICU-LOS. Given that bleeding patients tend to have longer ICU-LOS as determined by the causal framework, the regression model therefore targets a more homogeneous group of patients at risk of prolonged ICU-LOS compared to the overall heterogeneous ICU population.

Prediction results for patients undergoing non-surgical procedures from the Mayo Clinic perioperative datamart showed that the classification model can effectively identify patients at risk of all courses of prolonged ICU-LOS and bleeding in patients offered PPT with AUC as high as 0.84, sensitivity of 0.79 and G-mean of 0.75. The regression model on the other hand can identify patients at risk of prolonged ICU-LOS attributable to bleeding in patients offered PPT with acceptable accuracy (MAE=0.86, median predictions within 2 day of true median ICU-LOS). Through a simple transformation of the bleeding and PPT variables, the causal inference showed that the effect of bleeding in PPT patients significantly increases the risk of prolonged ICU-LOS (55%, p=0.00) and “no” bleeding in patients offered PPT significantly reduces ICU-LOS (4%, p=0.046).

Study Setting and Datasets

Over the past few years, there has been a significant increased in the application of machine learning methods to solve medical and health care problems. Often the problem reduces to applying standard classification or regression models. However, there are some problems that challenges standard application of machine learning methods. Two of such problems are considered in this study: to find a meaningful association between longer ICU-LOS and bleeding due to PPT before surgery, and to identify patients who are at elevated risk of bleeding due to PPT and those at risk of all causes of extended ICU-LOS. This section presents a causal and predictive framework in attempt to address these problems.

Outcome-Treatment Variable Transformation

To facilitate the estimation of the effect of bleeding in patients offered PPT on prolonged ICU-LOS and prediction of the risk of bleeding, the training data is grouped into 4 mutually exclusive classes as shown in Table 1. Class A represents patients who bled and were given PPT while class B are patients who did not bleed but were also given PPT. A similar interpretation can be made for the other classes. Using baseline covariates measured before PPT, causal and predictive models can be constructed to to make inference on the four classes. Specifically, the causal inference considers each of the classes in Table 1 as a treatment or exposure variable and then computes its effect on prolonged ICU-LOS. Thus estimates of the treatment effect in class A can be interpreted as the causal effect of bleeding as a “complication” of PPT on prolonged ICU-LOS. The predictive inference on the other hand simply constructs binary classification models for each of the four classes. A multiclass classification model can equally be constructed for the 4 classes.
Table 1: Outcome-Treatment Variable Transformation

<table>
<thead>
<tr>
<th>PPT</th>
<th>bleeding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>p(bleeding = 1</td>
</tr>
<tr>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>p(bleeding = 1</td>
</tr>
<tr>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>p(bleeding = 0</td>
</tr>
<tr>
<td>0</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>p(bleeding = 0</td>
</tr>
</tbody>
</table>

$p(bleeding = 1|X, PPT = 1)$ = probability of bleeding given patient was offered PPT

Effect of Bleeding as complication of PPT on prolonged ICU-LOS

The standard approach to investigate the causal relationship between a treatment or exposure (e.g. PPT or bleeding) and an outcome (e.g. ICU-LOS) is to construct statistical regression models in which the outcome is regressed against baseline covariates and the treatment variable. The attributable effect of the treatment is then read off as the corresponding regression coefficient. This study takes a different approach and estimates the treatment effect through application of machine learning methods. The theory of causal inference or technical details of the considered estimation procedure are beyond the scope of this study. The interested reader is referred to 13-15 for more details. However, for the purposes of this study, a brief discussion of the data structure required to compute these estimators is presented next.

Data structure and likelihood. The observations for each patient in the data set can be written as $O = (X, Y, Z)$ where $Z \in \{0, 1\}$ is the treatment indicator (e.g. class A, B, C or D in Table 1) with $Z = 1$ if patient was treated and $Z = 0$ if patient was not treated. $X$ is a vector of baseline covariates that records information specific to each patient prior to treatment. $Y$ is the outcome such as prolonged ICU-LOS (e.g. $Y = 1$ if ICU-LOS ≥ 7 days and $Y = 0$ if ICU-LOS < 7 days). The relationship between the observed variables in $O$ can be written in a factorize data likelihood as

$$Pr(X, Y, Z) = \frac{Q}{Pr(Z|X)Pr(Y|X, Z)Pr(X)}.$$  \hspace{1cm} (1)

$Pr(X)$ and $Pr(Y|X, Z)$ are referred to as the $Q$ component of the likelihood while $Pr(Z|X)$ is the $g$ component. $g(Z|X)$ represents the propensity or the causal disposition of the treatment to produce some outcome. $Q(Z, X) = E[Y|Z, X]$ is the expected potential outcomes conditional on the observed characteristics. Estimates of $g$ and $Q$ can be obtained by standard regression or machine learning methods.

For a binary outcomes and in the presence of no confounding variables, a straightforward approach to obtain the treatment effect is to compute the expectations $\psi_1 = E[Y|Z=1]$ and $\psi_0 = E[Y|Z=0]$, where $E[Y|Z=1]$ is the mean of $Y$ assuming every patient in the population was exposed at level $Z = 1$. These two statistics can then be combined in useful ways to assess the effect of different levels of the treatment. Three commonly reported summary statistics include: (1) Additive Treatment Effect : $ATE = \psi_1 - \psi_0$, (2) Risk Ratio : $RR = \psi_1/\psi_0$, and (3) Odds Ratio : $RO = RR \times (1 - \psi_0)/(1 - \psi_1)$.

The $ATE$ quantifies the additive effect of every patient being exposed to the event versus not being exposed. Thus if the event is class A i.e. bleeding as a complication of PPT, a meaningful interpretation of $ATE = 0.05$ could read: “being exposed to the event of bleeding as a complication of PPT versus not increases the risk of prolonged ICU-LOS by 5%”. The $RR$ quantifies the multiplicative effect of being exposed versus not. A $RR$ of 5 can be interpreted as: “being exposed to the event of bleeding as a complication of PPT versus not would lead to a 5 times increase in the risk of prolonged ICU-LOS”. Since the $OR$ is a function of $RR$, a similar interpretation can be given for $OR$.

Targeted maximum likelihood estimation. In observational studies, estimators of treatment effect need to account for possible confounding, i.e the (apparent) effect of the treatment is actually the effect of another characteristic which
is associated with the treatment and with the outcome. Several methods have been proposed for the estimation of ATE, RR and OR that can mitigate the effects of confounding (and model misspecification), e.g. G-computation formula, propensity score matching, inverse probability of treatment weighting (IPTW), and doubly-robust estimation. See 13,14 for more in-depth discussion of these estimators. In this study, the Targeted maximum likelihood estimation (TMLE) 15,16 method is considered because of its double robustness and bias reduction properties. TMLE is a two stage doubly robust semi-parametric estimation methodology designed to minimize the bias of the parameters of interest. The first stage of the method estimates the density of the data generating distribution (specifically $Q$) while the second stage solves an efficient influence curve estimating equation. The influence curve describes the behavior of the target parameter under slight changes of the initial density estimates. In TMLE, if either $g$ or $Q$ are consistently estimated, then the TMLE estimator is guaranteed to be asymptotically unbiased. However, TMLE will not return consistent estimates of the parameter of interest when both $g$ and $Q$ are misspecified. Thus it is important to avoid overfitting these measures.

As discussed above, estimating the two statistics $\psi_1$ and $\psi_0$ allows for calculating any of the causal effects ATE, RR and OR. The TMLE estimate of $\psi_z (z \in \{0, 1\})$ is given by

$$\hat{\psi}_z = \frac{1}{n} \sum_{i=1}^{n} Q^*(z, x_i)$$

where $Q^*(z, x_i)$ is an update of $Q(z, x_i)^2$. The targeting step for updating $Q(z, x_i)$ is done by fluctuating $Q(z, x_i)$ through a parametric sub-model of the form: $logit(Q^*(z, x)) = logit(Q(z, x)) + \varepsilon H_z(z, x)$, where $\varepsilon$ is the fluctuation parameter, $H(z, x) = I(Z = z)/\hat{g}(z, x)$ is the efficient influence curve equations, and $I$ is the indicator function. The MLE of $\varepsilon$ is obtained by a logistic regression of $Y$ on $H_z(z, x)$ with offset $logit(Q(z, x))$. Confidence intervals and p-value for TMLE can be obtained through the variance of the influence curve.

TMLE can use initial estimates of $Q$ and $g$ from any fixed parametric model such as generalized linear models (GLM) (e.g logistic regression). However, parametric models require assumptions regarding the functional form, distribution of variables and variable selection which are often not realistic such that model misspecification is difficult to avoid. It is therefore recommended to used machine learning methods that makes little or no assumptions and are able to estimate complex relationships between the outcome and observed variables.

**Two Level Predictive Model**

![Figure 1: Two Level Predictive Model](image)

In the proposed two level predictive model, classification models are constructed to predict each of the classes in Table 1. Specifically, a classifier is trained on a modified training set defined by $\{X, Y\}$, where $X$ is a matrix containing the patients baseline covariates and $Y$ is a binary variable representing one of the classes in Table 1 such that $Y = 1$ for all

---

2the hat (‘’) notation represents estimates of a parameter from the data
patients in the corresponding class and 0 otherwise. During training, the classifier is trained on the complete training set to predict $Y$ and a regression model is trained to predict ICU-LOS using a subset of the training set for which $Y = 1$. In the case where class A is the outcome of interest, the second level regression model is trained to specifically target patient at high risk of longer ICU-LOS. This is as a consequence of the known relationship between PPT, bleeding, and prolonged ICU-LOS from the literature $^{3,4,12}$, and specifically from causal relationship between these variables derived in this study.

Model validation proceeds as follows; given a new patient, the classification model is first applied to predict the outcome $\hat{Y}$ for the patient. If $\hat{Y} = 1$, then the regression model is applied to predict the corresponding ICU-LOS. Figure 1 illustrates the training and testing steps of the two level predictive model.

**Predicting prolonged ICU-LOS attributable to Bleeding as a complication of PPT.** Given that class A in Table 1 represents patients who bled and were administered PPT and the fact that bleeding has been found to be significantly associated with prolonged ICU-LOS $^{3,4,12}$, a regression model to predict ICU-LOS is constructed only for patients in this sub-population. The rationale behind this strategy is that since this sub-population is relatively homogeneous, the approach allows for optimal targeting of patients at risk of longer ICU-LOS attributable to bleeding as a complication of PPT compared to a model constructed on the general ICU population. Similar regression models can also be constructed for patients identified in each of the other classes in Table 1 to predict the actual ICU-LOS. Class B is of particular interest as these patients did not bleed but were administered PPT. The study of these sub-groups with respect to prolonged ICU-LOS could potentially shed lights on the characteristics of patients who stand to benefit from plasma transfusions.

**Predicting all causes of prolonged ICU-LOS.** Without taking into account any particular cause, prolonged ICU-LOS was defined as ICU-LOS $\geq 7$ days and classification models were constructed to predict this binary outcome. The 7 days threshold was chosen because it represented a reasonable number of days where physicians might be concern about a prolonged stay. This variable was also used in TMLE estimation to estimate the effects of the four exposures in Table 1 on ICU stays of 7 days or more.

**Experiments**

This section describes the study population and presents empirical evaluation results.

**Study Population**

The data used in this study are derived from the Mayo Clinic transfusion datamart$^{17}$ that captures demographics, disease conditions, laboratory test results, medications, operative and postoperative measurements and outcomes for all patients admitted to acute care environments. To be considered for study participation, patients must meet the following criteria: age $\geq 18$ years, noncardiac surgery and an INR $\geq 1.5$ in the 30 days preceding surgery. Between 2008 and 2011, a total of 1,233 patients were identified and comprised the study population.

**Baseline Variables.** Baseline patient demographics include age, height, weight, gender and the ASA status I-V classifications. Disease conditions included myocardial infarction, congestive heart failure, cerebrovascular disease, dementia, chronic pulmonary disease, diabetes mellitus, etc. Preoperative laboratory values included INR, hemoglobin, platelet counts, creatinine, albumin, and APTT. A total of 51 predictors were considered for inclusion in the analyses. Table 2 presents the characteristics segmented according to the four classes in Table 1. The table show that the four groups are significantly different across demographics (except gender), laboratory, medications, and planned procedure types. The groups are however similar across majority of disease conditions. Further, patients who bled and were offered PPT had significantly longer ICU-LOS compared to those who did not.

**Outcomes and Treatment.** Bleeding was taken as the World Health Organization (WHO) grade 3 bleeding events, defined as the need for early perioperative RBC transfusion$^{18}$. The term perioperative RBC transfusion was defined as the administration of allogeneic RBC components during the interval beginning with entry into the operating room and ending 24 hours after exit from the operating room. ICU-LOS is the number of days spent in the ICU after surgery. The treatment indicator PPT indicates if a patient was offered plasma transfusion after INR test and before surgery. ICU-LOS
Table 2: Patient characteristics by classes in Table 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>&lt; 61</td>
<td>≥ 61</td>
<td>&lt; 61</td>
<td>≥ 61</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Weight</td>
<td>&lt; 85</td>
<td>≥ 85</td>
<td>&lt; 85</td>
<td>≥ 85</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Gender: M</td>
<td>62%</td>
<td>64%</td>
<td>59%</td>
<td>60%</td>
<td>0.92</td>
</tr>
<tr>
<td>ASA Physical Status</td>
<td>2</td>
<td>1</td>
<td>12%</td>
<td>≥ 12%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>33%</td>
<td>74%</td>
<td>46%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>52%</td>
<td>14%</td>
<td>46%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>12%</td>
<td>0%</td>
<td>3%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Disease conditions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>23%</td>
<td>32%</td>
<td>27%</td>
<td>28%</td>
<td>0.71</td>
</tr>
<tr>
<td>Cerebrovascular Disease</td>
<td>22%</td>
<td>23%</td>
<td>14%</td>
<td>23%</td>
<td>0.000</td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>36%</td>
<td>41%</td>
<td>26%</td>
<td>36%</td>
<td>0.002</td>
</tr>
<tr>
<td>Pulmonary Disease</td>
<td>18%</td>
<td>17%</td>
<td>12%</td>
<td>12%</td>
<td>0.33</td>
</tr>
<tr>
<td>Chronic Renal Failure</td>
<td>32%</td>
<td>17%</td>
<td>23%</td>
<td>18%</td>
<td>0.011</td>
</tr>
<tr>
<td>Dementia</td>
<td>0%</td>
<td>5%</td>
<td>3%</td>
<td>2%</td>
<td>0.18</td>
</tr>
<tr>
<td>Diabetes</td>
<td>40%</td>
<td>30%</td>
<td>35%</td>
<td>36%</td>
<td>0.71</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>0.6</td>
</tr>
<tr>
<td>Connective Tissue Disease</td>
<td>10%</td>
<td>5%</td>
<td>3%</td>
<td>5%</td>
<td>0.001</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>5%</td>
<td>2%</td>
<td>3%</td>
<td>3%</td>
<td>0.56</td>
</tr>
<tr>
<td>MI</td>
<td>19%</td>
<td>15%</td>
<td>12%</td>
<td>15%</td>
<td>0.45</td>
</tr>
<tr>
<td>Peptic Ulcer</td>
<td>4%</td>
<td>5%</td>
<td>3%</td>
<td>5%</td>
<td>0.82</td>
</tr>
<tr>
<td>Peripheral Vascular Disease</td>
<td>5%</td>
<td>3%</td>
<td>3%</td>
<td>5%</td>
<td>0.42</td>
</tr>
<tr>
<td>Mild Liver Disease</td>
<td>18%</td>
<td>9%</td>
<td>37%</td>
<td>8%</td>
<td>0.001</td>
</tr>
<tr>
<td>Moderate/Severe Liver Disease</td>
<td>1%</td>
<td>2%</td>
<td>20%</td>
<td>3%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Charlson score</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>&lt; 0.011</td>
</tr>
<tr>
<td><strong>Laboratory test results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INR</td>
<td>1.8</td>
<td>1.7</td>
<td>1.8</td>
<td>&gt; 1.7</td>
<td>&lt; 0.045</td>
</tr>
<tr>
<td>Platelets</td>
<td>11%</td>
<td>3%</td>
<td>3%</td>
<td>1%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>APIT</td>
<td>167</td>
<td>134</td>
<td>129</td>
<td>&lt; 205</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Creatinine</td>
<td>1.4</td>
<td>1.1</td>
<td>1.2</td>
<td>&gt; 1.0</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>10.1</td>
<td>&lt; 12.9</td>
<td>&lt; 9.8</td>
<td>&lt; 12.3</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>PLT</td>
<td>29%</td>
<td>14%</td>
<td>41%</td>
<td>8%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Medications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspirin</td>
<td>68%</td>
<td>53%</td>
<td>62%</td>
<td>52%</td>
<td>0.002</td>
</tr>
<tr>
<td>Clopidogrel</td>
<td>8%</td>
<td>8%</td>
<td>3%</td>
<td>4%</td>
<td>0.053</td>
</tr>
<tr>
<td>Coumadin</td>
<td>55%</td>
<td>71%</td>
<td>43%</td>
<td>60%</td>
<td>0.001</td>
</tr>
<tr>
<td>Heparin</td>
<td>36%</td>
<td>24%</td>
<td>20%</td>
<td>19%</td>
<td>0.006</td>
</tr>
<tr>
<td><strong>Planned Procedure</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENT/Otolary</td>
<td>0%</td>
<td>6%</td>
<td>4%</td>
<td>10%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>General</td>
<td>58%</td>
<td>59%</td>
<td>25%</td>
<td>28%</td>
<td>0.001</td>
</tr>
<tr>
<td>Neurology</td>
<td>3%</td>
<td>3%</td>
<td>1%</td>
<td>2%</td>
<td>0.77</td>
</tr>
<tr>
<td>O/G</td>
<td>1%</td>
<td>2%</td>
<td>1%</td>
<td>4%</td>
<td>0.076</td>
</tr>
<tr>
<td>Orthopedic</td>
<td>4%</td>
<td>5%</td>
<td>3%</td>
<td>2%</td>
<td>0.001</td>
</tr>
<tr>
<td>Thoracic</td>
<td>10%</td>
<td>5%</td>
<td>4%</td>
<td>5%</td>
<td>0.27</td>
</tr>
<tr>
<td>Transplantation</td>
<td>19%</td>
<td>2%</td>
<td>37%</td>
<td>2%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Urology</td>
<td>1%</td>
<td>3%</td>
<td>2%</td>
<td>9%</td>
<td>0.001</td>
</tr>
<tr>
<td>Vascular</td>
<td>5%</td>
<td>9%</td>
<td>6%</td>
<td>5%</td>
<td>0.29</td>
</tr>
<tr>
<td>Emergency</td>
<td>63%</td>
<td>38%</td>
<td>48%</td>
<td>12%</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

ICU-LOS: A: 5.1, B: > 6.0, C: > 7.1, D: > 8.0, p < 0.001

Results

To demonstrate the effectiveness of the two level approach in identifying patients at risk of longer ICU-LOS attributable to bleeding as a complication of PPT, the second level regression model is compared with a regression model trained on the complete training data. The median predicted ICU-LOS and the true median ICU-LOS are also reported for the complete test set and for the top 75% percentile of the true distribution of ICU-LOS.

Figure 2 shows the distribution of observed ICU-LOS for the entire population truncated at 250 patients with no ICU admission and 30 days length of stay. It can be seen that the distribution of ICU stays is highly skewed to the right. Learning to identify extreme values at the tail of the distribution can be a very challenging task for a regression model constructed using the entire data. However, this task can be made easier if one focuses only on the extreme values.
The random forest (RF) and the gradient boosting machine (GBM) algorithms were considered for classification and regression. The algorithms were trained and evaluated using 100 bootstraps. On each bootstrap iterate, the algorithms are trained on approximately 63% of the data and the left out samples was used for testing. Model selection, i.e. hyperparameter tuning (by grid search) was done using a 5-fold cross-validation procedure. Random forest was trained with 1000 trees.

In calculating estimates of TMLE, RF and GBM along with 6 other algorithms: logistic regression, k-nearest neighbors, support vector machines, neural networks, decision trees, and lasso and elastic net were combined in a Super Learner algorithm. Super Learning is a strategy for combining several data-adaptive estimators into one improved estimator. Specifically, the functions $g$ and $Q_z$ in equation (2) are each estimated by these algorithms and combined by the Super Learner to produce better estimates. For comparison, estimates of TMLE are also computed using standard logistic regression (GLM).

**Table 3: Effect of bleeding as a complication of PPT on prolonged ICU-LOS**

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Super Learner</th>
<th>GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>p.value</td>
</tr>
<tr>
<td>A</td>
<td>ATE</td>
<td>0.57 (0.55,0.59)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>8.43 (6.97,10.20)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>22.17 (17.95,27.37)</td>
</tr>
<tr>
<td>B</td>
<td>ATE</td>
<td>-0.04 (-0.08,0.00)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>0.52 (0.22,1.26)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>0.50 (0.20,1.26)</td>
</tr>
<tr>
<td>C</td>
<td>ATE</td>
<td>0.09 (0.07,0.11)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>2.08 (1.71,2.53)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>2.31 (1.86,2.88)</td>
</tr>
<tr>
<td>D</td>
<td>ATE</td>
<td>-0.05 (-0.05,-0.01)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>0.69 (0.56,0.86)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>0.67 (0.55,0.85)</td>
</tr>
<tr>
<td>bleeding</td>
<td>ATE</td>
<td>0.17 (0.15,0.20)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>3.57 (2.86,4.44)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>4.38 (3.44,5.58)</td>
</tr>
<tr>
<td>PPT</td>
<td>ATE</td>
<td>0.04 (0.04,0.08)</td>
</tr>
<tr>
<td></td>
<td>RR</td>
<td>1.73 (1.40,2.147)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>1.85 (1.45,2.35)</td>
</tr>
</tbody>
</table>

**Effect of Bleeding as a complication of PPT on prolonged ICU-LOS.** Table 3 presents the TMLE estimates of ATE, RR, and OR quantifying the effect of each of the exposures in Table 1, all causes of bleeding, and PPT on prolonged ICU-LOS. Results for TMLE using the Super Learner and a simple main effect GLM are shown for comparison. However, it should be noted that diagnostic tests to ensure that the GLM model fit the data were not performed, thus subsequent discussions will be based on the Super Learner. To save on space, the discussions will also be restricted to
the ATE summary statistics; interpretations for RR and OR can be similarly made.

Results in Table 3 show that bleeding as a complication of PPT (exposure A in Table 1), all causes of bleeding, and PPT significantly increases the risk of prolonged ICU-LOS by 57%, 17% and 4% respectively (95% confidence intervals are shown in brackets). The event of no bleeding in patients offered PPT (exposure B) was found to reduce the risk of prolonged ICU by 4% (p-value = 0.05) while the event of bleeding when no PPT was administered (exposure C) increases the risk of prolonged ICU-LOS by 9%. These results illustrates that by a simple transformation of the treatment variable, one can identify groups of patients with differential effects of the treatment. Patients identified in group B might show beneficial effects of PPT and thus can be targeted for treatment to mitigate bleeding and reduce ICU-LOS.

Table 4: Performance of classification models

<table>
<thead>
<tr>
<th>Model</th>
<th>Class</th>
<th>AUC</th>
<th>sensitivity</th>
<th>G.mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>A</td>
<td>0.82 (0.76,0.88)</td>
<td>0.67 (0.20,0.94)</td>
<td>0.72 (0.43,0.79)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.67 (0.55,0.79)</td>
<td>0.50 (0.06,0.83)</td>
<td>0.58 (0.25,0.70)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.83 (0.77,0.88)</td>
<td>0.72 (0.56,0.85)</td>
<td>0.74 (0.69,0.79)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.84 (0.79,0.86)</td>
<td>0.78 (0.59,0.89)</td>
<td>0.75 (0.70,0.79)</td>
</tr>
<tr>
<td>GBM</td>
<td>A</td>
<td>0.80 (0.72,0.88)</td>
<td>0.65 (0.19,0.95)</td>
<td>0.70 (0.41,0.81)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.65 (0.51,0.73)</td>
<td>0.48 (0.13,0.75)</td>
<td>0.58 (0.35,0.68)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.83 (0.80,0.87)</td>
<td>0.69 (0.56,0.87)</td>
<td>0.74 (0.60,0.78)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.83 (0.80,0.86)</td>
<td>0.79 (0.66,0.89)</td>
<td>0.74 (0.70,0.78)</td>
</tr>
<tr>
<td>Multiclass RF</td>
<td>A</td>
<td>0.78 (0.75,0.81)</td>
<td>0.38 (0.35,0.42)</td>
<td>0.57 (0.54,0.60)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.80 (0.79,0.80)</td>
<td>0.69 (0.68,0.74)</td>
<td>0.74 (0.74,0.74)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.79 (0.73,0.79)</td>
<td>0.70 (0.45,0.72)</td>
<td>0.70 (0.61,0.72)</td>
</tr>
</tbody>
</table>

Performance of Classification Models. Table 4 presents performance results for predicting the four classes in Table 1. Overall, the models performed very well. Best results for all classes was obtained for RF with AUC as high as 0.84 (class A) and 0.67 (class B). However, no significant difference was observed in the performances. For completeness, Table 4 also show the average per-class performance of a multiclass RF model trained using all four classes in Table 1. The the sensitivity of the multiclass model is however inferior to that of the least performing binary RF model.

Table 5: Performance of regression models for predicting log(ICU-LOS)

<table>
<thead>
<tr>
<th>Model</th>
<th>Class</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>A</td>
<td>0.86 (0.59,1.04)</td>
<td>1.04 (0.73,1.19)</td>
<td>1.00 (0.77,1.36)</td>
<td>1.16 (0.99,1.51)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.71 (0.53,1.01)</td>
<td>0.88 (0.61,1.30)</td>
<td>0.64 (0.42,0.86)</td>
<td>0.81 (0.53,1.13)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.68 (0.61,0.76)</td>
<td>0.85 (0.77,0.96)</td>
<td>0.66 (0.57,0.78)</td>
<td>0.85 (0.74,1.02)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.44 (0.39,0.48)</td>
<td>0.68 (0.59,0.77)</td>
<td>0.48 (0.44,0.54)</td>
<td>0.68 (0.61,0.79)</td>
</tr>
<tr>
<td>GBM</td>
<td>A</td>
<td>0.88 (0.66,1.01)</td>
<td>1.05 (0.80,1.19)</td>
<td>0.99 (0.80,1.23)</td>
<td>1.18 (0.97,1.42)</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.69 (0.42,0.97)</td>
<td>0.85 (0.56,1.30)</td>
<td>0.64 (0.41,0.87)</td>
<td>0.79 (0.58,1.13)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.70 (0.61,0.77)</td>
<td>0.87 (0.77,0.97)</td>
<td>0.67 (0.60,0.81)</td>
<td>0.86 (0.75,1.07)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.42 (0.37,0.47)</td>
<td>0.70 (0.61,0.78)</td>
<td>0.46 (0.41,0.52)</td>
<td>0.69 (0.60,0.77)</td>
</tr>
</tbody>
</table>

Performance of Regression Models. Results for the two level predictive model (TwoLevel) as previously described is compared with regression models trained on the complete training set (OneLevel). Table 5 shows that the TwoLevel training approach makes smaller errors for classes A and D while OneLevel training makes smaller errors for classes B and C. These differences were however not significant. A close look at the actual predicted median ICU-LOS in days (see Table 6) for the top 75% percentile of the true distribution of ICU-LOS (i.e. longer ICU-LOS) show that TwoLevel regression models tend to make smaller errors for all classes. Overall, the predicted median ICU length of stay from the TwoLevel model were off from the observed median value by less than 2 days while OneLevel was at least 4 days.
Table 6: Predicted and observed median ICU-LOS (in days)

<table>
<thead>
<tr>
<th>Model</th>
<th>Class</th>
<th>TwoLevel</th>
<th>OneLevel</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>3.99 (2.55,5.36)</td>
<td>1.86 (1.00,2.42)</td>
<td>5.00 (2.18,6.66)</td>
</tr>
<tr>
<td>RF</td>
<td>B</td>
<td>0.83 (0.48,1.40)</td>
<td>0.70 (0.41,1.33)</td>
<td>0.00 (0.00,0.98)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1.36 (1.12,1.60)</td>
<td>1.19 (0.95,1.37)</td>
<td>1.15 (0.93,1.73)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.26 (0.20,0.36)</td>
<td>0.37 (0.28,0.46)</td>
<td>0.00 (0.00,0.00)</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>4.05 (2.70,5.90)</td>
<td>1.78 (1.08,3.15)</td>
<td>5.14 (2.26,7.01)</td>
</tr>
<tr>
<td>GBM</td>
<td>B</td>
<td>0.69 (0.47,1.31)</td>
<td>0.71 (0.37,1.39)</td>
<td>0.00 (0.00,0.81)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1.31 (1.05,1.53)</td>
<td>1.07 (0.75,1.29)</td>
<td>1.11 (0.83,1.60)</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>0.22 (0.14,0.31)</td>
<td>0.32 (0.25,0.44)</td>
<td>0.00 (0.00,0.00)</td>
</tr>
</tbody>
</table>

Top 75% predicted and observed median ICU-LOS

A 12.08 (8.67,37.04) 9.01 (8.44,9.57) 15.19 (8.41,47.63)
B 2.91 (1.85,9.54) 2.65 (1.84,4.53) 4.37 (1.79,50.33)
C 3.97 (2.95,9.84) 3.59 (2.96,7.08) 5.99 (2.93,37.02)
D 0.33 (0.02,1.94) 0.48 (0.03,2.72) 0.00 (0.00,12.96)

These results show that using only pretreatment variables, with the TwoLevel model one can predict with high accuracy the risk of a patient bleeding as a complication of PPT and then use this information to determine if the patient is likely going to stay longer in the ICU with sufficient accuracy.

Important Characteristics of Patients at Risk of Prolonged ICU-LOS. Table 7 presents the top 10 important variables for the RF regression model (averaged over 100 bootstrap). The TwoLevel model identified the INR for patients who bled and PLT (platelet count) for patients who did not bleed in the case where these patients were offered PPT as the most predictive variables. Hemoglobin level was the most predictive characteristics for all patients. Though appearing in different order, six top variables are common to all models.

Table 7: Random Forest Regression Model Top 10 Variable Importance

<table>
<thead>
<tr>
<th>Variables</th>
<th>Class A</th>
<th>Class B</th>
<th>OneLevel Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>INR</td>
<td>85.6 (60.05,100.00)</td>
<td>95.67 (82.79,100.00)</td>
<td>Hemoglobin 99.34 (95.02,100.00)</td>
</tr>
<tr>
<td>PLT</td>
<td>77.5 (46.61,100.00)</td>
<td>91.95 (78.81,100.00)</td>
<td>Emergency 94.29 (83.83,100.00)</td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>77.35 (36.50,100.00)</td>
<td>89.97 (70.63,100.00)</td>
<td>PLT 94.22 (85.37,100.00)</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>72.85 (27.30,100.00)</td>
<td>87.79 (72.18,100.00)</td>
<td>ASA:4 90.13 (82.33,97.60)</td>
</tr>
<tr>
<td>Age</td>
<td>70.91 (42.76,93.12)</td>
<td>87.32 (72.42,98.06)</td>
<td>Age 87.96 (79.08,96.73)</td>
</tr>
<tr>
<td>Creatinine</td>
<td>70.2 (40.70,100.00)</td>
<td>85.1 (71.45,98.74)</td>
<td>Creatinine 87.87 (77.44,96.74)</td>
</tr>
<tr>
<td>Procedure:ENT/Oral</td>
<td>65.13 (0.10,100.00)</td>
<td>85.89 (71.70,96.97)</td>
<td>Weight 85.29 (75.03,94.53)</td>
</tr>
<tr>
<td>Emergency</td>
<td>65.02 (30.75,100.00)</td>
<td>82.2 (70.21,96.69)</td>
<td>INR 80.62 (70.96,89.70)</td>
</tr>
<tr>
<td>Weight</td>
<td>62.97 (30.42,97.50)</td>
<td>81.65 (67.06,98.39)</td>
<td>Charlson score 80.34 (70.47,90.24)</td>
</tr>
<tr>
<td>ASA:3</td>
<td>61.34 (28.24,92.29)</td>
<td>80.92 (63.92,99.55)</td>
<td>Heparin 77.45 (59.86,89.27)</td>
</tr>
</tbody>
</table>

Conclusion

This study establishes a previously unknown causal relationship between prolonged ICU-LOS, bleeding, and PPT. Through a simple transformation of the bleeding and PPT variables, it was shown that the effect of bleeding in patients offered PPT more than doubled the risk of prolonged ICU-LOS, and that the event of no bleeding in patients also offered PPT significantly reduces the risk of prolonged ICU-LOS. Motivated by these results, this study developed a predictive framework for early identification of patients with (1) increased risk of all causes of prolonged ICU-LOS, and prolonged ICU-LOS attributable to bleeding as a complication of PPT, and (2) increased risk of bleeding as a complication of PPT. Given that the performance of models for ICU-LOS often deteriorate for patients with extended ICU-LOS, a two level predictive modeling approach was proposed to improve performance at the tail of the ICU-LOS distribution. The two level model was shown through a series of experiments using data for patients undergoing non-surgical procedures to be capable of identifying patient at risk or bleeding and longer ICU-LOS with acceptable accuracy. Based on the characteristics and predicted risk scores generated by the framework, patients at risk of extended ICU stays can be targeted and special intervention procedures implemented to reduce ICU-LOS and
improve perioperative care.

A possible weakness of this study is the use of the treatment effects \textit{ATE, RR} and \textit{OR}. As these are population averages, their use entails that the effect is uniform across sub-populations or individuals and therefore not very useful for characterizing the effect of the treatment for each patient. Thus, an interesting potential future study of this work is to investigate methods of combining the four probabilities in Table 1 to derive individualized treatment effects.

References

Automated Cancer Registry Notifications: Validation of a Medical Text Analytics System for Identifying Patients with Cancer from a State-Wide Pathology Repository

Anthony N. Nguyen, PhD¹, Julie Moore², John O’Dwyer¹, Shoni Philpot²
¹The Australian e-Health Research Centre, CSIRO, Brisbane, Australia;
²Queensland Cancer Control Analysis Team, Department of Health, Queensland Government, Brisbane, Australia

Abstract

The paper assesses the utility of Medtex on automating Cancer Registry notifications from narrative histology and cytology reports from the Queensland state-wide pathology information system. A corpus of 45.3 million pathology HL7 messages (including 119,581 histology and cytology reports) from a Queensland pathology repository for the year of 2009 was analysed by Medtex for cancer notification. Reports analysed by Medtex were consolidated at a patient level and compared against patients with notifiable cancers from the Queensland Oncology Repository (QOR). A stratified random sample of 1,000 patients was manually reviewed by a cancer clinical coder to analyse agreements and discrepancies. Sensitivity of 96.5% (95% confidence interval: 94.5-97.8%), specificity of 96.5% (95.3-97.4%) and positive predictive value of 83.7% (79.6-86.8%) were achieved for identifying cancer notifiable patients. Medtex achieved high sensitivity and specificity across the breadth of cancers, report types, pathology laboratories and pathologists throughout the State of Queensland. The high sensitivity also resulted in the identification of cancer patients that were not found in the QOR. High sensitivity was at the expense of positive predictive value; however, these cases may be considered as lower priority to Cancer Registries as they can be quickly reviewed. Error analysis revealed that system errors tended to be tumour stream dependent. Medtex is proving to be a promising medical text analytic system. High value cancer information can be generated through intelligent data classification and extraction on large volumes of unstructured pathology reports.

Introduction

The Queensland Cancer Registry (QCR) is a population based cancer registry that monitors and records the incidence and mortality of cancer in the State of Queensland, Australia over time by collecting cancer notifications from a variety of sources. The QCR collection, therefore, has a number of intrinsic values, which it derives from its population base, and provides the capacity to support longitudinal analysis. The QCR also supports key activities in Queensland and nationally such as health service planning and cancer research.

Pathology laboratories throughout Queensland are legally required to notify the Cancer Registry of pathology tests that contain a result of cancer. In Queensland, notifiable cancers to the registry include:

1. All invasive cancers excluding basal cell carcinoma (BCC) and squamous cell carcinomas (SCC) of the skin;
2. Any cancer with uncertain behaviour;
3. All in-situ conditions; and

All of the pathology cancer notifications are currently paper based and the incidence of cancers is increasing with the number of new cancer cases in Queensland increasing by more than 177% between 1982 and 2012; the growth in new cancer cases is largely due to population growth and ageing¹. A growing backlog is delaying the delivery of more timely cancer information due to the extent of manual processing and an out-dated information collection system. This highlights the need for introducing new technologies to assist with automating cancer registry processes.

With some updated technology changes, pathology laboratories can now send pathology electronically via HL7 feeds to the Queensland Oncology Repository (QOR). However, this is still not without its challenges. There is no mechanism to determine whether an electronic pathology message is cancer notifiable or not, apart from reading and interpreting the contents of the HL7 pathology report. As such, a computer-assisted approach is required to automatically identify pathology reports that are cancer notifiable.
Background

A number of systems have been proposed to address the automatic detection of cancer notifiable (or reportable) pathology reports. These systems rely solely on the use of custom made list of cancer and non-cancer related terms, phrases, and disease codes that may be institution specific, or the development of reportable cancer or tumour specific machine learning classification models\(^1\)\(^-\)\(^5\). The proposed method however aims to be generic supporting the full range of tumour streams (or types of cancers) from a breadth of pathology laboratories, report types and pathologists.

Commercial cancer finding and reporting systems that selects reports that contain reportable cancer findings based on dictionary and linguistic analysis are also available\(^6\)\(^-\)\(^7\). Although, conceptually similar to the proposed approach, it is unclear as to how much modification is required to achieve high levels of accuracy within the context of an Australian Cancer Registry.

The proposed approach is based on Medtex\(^8\). Medtex is an emerging medical text analytic capability that conducts intelligent data classification and extraction on large volumes of unstructured pathology reports, which can result in the generation of high value cancer information. Medtex has been developed to automate Cancer Registry notification using QCR business rules, natural language processing, and symbolic reasoning using SNOMED CT subsumption querying\(^9\)\(^-\)\(^13\). Medtex can assist in supporting the continuous improvement of cancer notification and enables improved decision support for Registry clinical coders.

Related work on automating cancer notifications using Medtex has shown that the system classified cancer notifiable reports with a sensitivity, specificity, and positive predictive value (PPV) of 0.98, 0.96, and 0.96, respectively, for an evaluation set of 479 histology and cytology reports\(^9\). Although very promising results were achieved, outperforming many of the systems in the literature, the dataset was very limited in size. The reports were selectively sampled such that an approximately balanced dataset resulted with half of the dataset covering a range of cancers, while the other half containing non-cancers. In spite of this, Medtex aims to be generic and was developed using Cancer Registry business rules and the normalisation and reasoning of reports using standard clinical terminologies; it does not rely on custom phrases, explicitly mentioned disease codes, or the development of machine-learning classification models\(^2\)\(^-\)\(^5\).

This research extends previous work and evaluates the utility of Medtex to identify notifiable cancer pathology reports on a larger and representative cancer population based dataset. Results show that high sensitivity and specificity can be achieved for a representative dataset containing a breadth of tumour streams, report types, pathology laboratories and pathologists. Analysis of discrepancies revealed that Medtex was able to 1) identify cancer patients not found in the QOR potentially underestimating the incidence of cancer and 2) actual system errors tended to be tumour stream dependent. These insights were not possible with the previous study that used a smaller and unrepresentative dataset. These insights also allow for specific system limitations to be addressed in future developments.

Method

System description

A high-level architecture of the Medtex medical text analytic system for cancer notifications classification and data extraction is shown in Figure 1.

In this paper, the cancer notification classification component of the system will be evaluated. A two-pass approach is adopted to classify whether pathology reports were cancer notifiable or not\(^9\). The first pass queries and filters pathology HL7 messages from the pathology data repository to identify report types that are required by the Cancer Registry, while the second pass analyses the free text reports and classifies them as either “Cancer Notifiable” (i.e., positive histology or cytology reports, excluding urine, sputum and pap smears), “Not Cancer Notifiable” or “Supporting Notifiable Reports” (i.e., those with further excisions resulting in no residual cancer, re-excisions and/or suspected notifiable cancers).

More specifically, Medtex is developed using the General Architecture for Text Engineering (GATE) platform\(^14\). It unifies the language in pathology reports by mapping spans in the text (e.g., clinical terms, abbreviations and acronyms, short-hand terms and relevant legacy codes) into SNOMED CT concepts\(^4\). Cancer notification identification is then based on finding the most likely primary site and histological type of the cancer based on SNOMED CT subsumption querying, and analyzing the report substructure and contextual information surrounding the concept to ensure that the concept’s mention was not modified by negative or uncertain assertions. Following QCR business rules, histological types that were either squamous or basal cell carcinoma (SCC or BCC) and was associated with the skin were classified as “Not Cancer Notifiable”\(^10\). Here, if SCC or BCC concepts were associated with a “skin” concept (i.e., concepts co-occurred within the same sentence), then such a classification would be made. The
QCR business rules are consistent and align with those nationally for the notification of cancers. More details on the cancer notification classification component of the system can be found in previous work.\(^9\)

The architecture supporting Medtex is characterised by a messaging framework built on the concept of message queues, producers and consumers\(^11,15\). Because multiple message consumers can be set up in parallel to receive messages from the same queue, Medtex provides high throughput, making it an ideal framework for analysing large streams of electronic pathology feeds.

Figure 1. High-level architecture of cancer notification classification and data extraction modules in Medtex.

The producer queries for pathology report types, which are potentially notifiable to the Cancer Registry (i.e., histology and cytology), and adds them to the input queue for Medtex processing. Medtex can then be set up either as a single instance or in parallel as consumers acting on the input queue, where each consumer will take a message from the input queue in turn for processing and analysis. The results from the Medtex analysis are then published to an output queue where another consumer can subscribe to the output queue to store the results in a SQL database. Figure 2 shows the database output schema portraying the range of cancer notification information extracted from pathology reports such as primary site of cancer, histology, grade, laterality, metastatic sites, as well as synoptic data including stage\(^9,13\).
Data collection and processing

With research ethics approval from the Queensland Health Research Ethics Committee, a corpus of 45.3 million HL7 pathology messages from the Queensland Oncology Repository (QOR), for the year of 2009 was used. QOR compiles and collates data from a range of Queensland cancer data source systems including the Queensland Cancer Registry (QCR), hospital admissions data, death data, treatment systems, public and private pathology and various hospital clinical data systems. The QOR is a resource managed by the Queensland Cancer Control Analysis Team (QCCAT) within the Department of Health, Queensland.

The corpus comprises of pathology laboratory results in the State of Queensland from the public health sector; these include bloods, tumour markers, etc. A total of 100 out of 114 possible unique report codes (or report types) were identified in the corpus. Upon filtering of the HL7 pathology feeds for histology and cytology report types (16 report types in total; e.g., histology frozen, histology biopsy, cytology (fluids), flow cytology, haematology) that are potentially relevant for Cancer Registry notifications, 119,581 histology and cytology reports remained for Medtex processing.

The histology and cytology reports were reported from 34 pathology laboratories in Queensland and therefore cover a breadth of report types, tumour streams (including non-cancer cases), laboratories and pathologists.

Performance evaluation

The 119,581 histology and cytology reports were analysed as a batch process by Medtex for reports requiring notification to the Cancer Registry.

These reports were consolidated at a patient level based on the exact matching of their first and last names, sex and date of birth (DOB). Patients with invalid or potentially multiple first/last name, sex and DOB combinations were removed from the analysis; for example, patients with punctuations in names (e.g., commas, quotes, hyphens), unknown sex and DOB. Patients with a combination of the above, for example, patients having same first name, sex and DOB but a different last name due to a hyphen (e.g., “FirstName LastName” and “FirstName MaidenName-LastName” with same DOB and Sex, and patients having the same first and last name and DOB as those with an unknown sex), were also removed from the analysis. The removal of such patients was to ensure that that the full set of reports processed by Medtex for a patient most likely corresponded to the set of reports for patients used by the Cancer Registry for cancer notifications coding.

A patient was classified as “Cancer Notifiable” if Medtex classified at least one of its reports as “Cancer Notifiable”, and vice versa, if Medtex classified all reports for a patient as “Not Cancer Notifiable”, then the patient was classified as “Not Cancer Notifiable”. For the purposes of the evaluations conducted in this paper, “Supporting Notifiable Report” classifications by Medtex were assumed “Cancer Notifiable” for the patient-level consolidation process. This assumption is likely to over estimate the number of patient-level cancer cases resulting from Medtex.
The QOR was used as the ground truth to identify actual patients diagnosed with cancer. A patient was considered to be “Cancer Notifiable” if the patient existed in the QOR database, and vice versa, the patient was assumed “Not Cancer Notifiable” if the patient could not be found in the QOR. A cancer clinical coder reviewed a stratified sample of 1,000 patients (each with potentially multiple reports) to analyse the agreements and discrepancies between Medtex and the QOR. A larger number of patients were selected from the stratum where Medtex and the QOR were in disagreement. This was purposely done to study the potential limitations of the system. The cancer clinical coder manually reviewed all the reports pertaining to the patients independently and classified their findings as “Cancer Notifiable”, “Not Cancer Notifiable” and “Supporting Notifiable Reports” in a similar way that Medtex classifies them. Again, results from a report-level were consolidated at a patient-level for analysis. Discrepancies between the coder’s judgements and that from Medtex and the QOR were analysed to uncover the cause of the discrepancy.

An adjusted contingency table (or confusion matrix) of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) was generated (in light of the review of discrepancies) to compute the sensitivity (or recall), specificity, positive predictive value (or precision) and F1 score of Medtex in classifying cancer notifications.

**Results**

Medtex filtered the 45.3 million HL7 pathology messages and processed and analysed the 119,581 histology and cytology reports in just under 5 days using 3 Medtex instances running in parallel (average processing rate of 3.6 seconds per report). A range of cancer characteristics were extracted, however, the current evaluation is only concerned with the cancer notification classification output by Medtex. Upon patient consolidation and the removal of 1,359 patients whose reports could not be reliably aggregated at a patient-level based on the exact match of first and last name, sex and DOB, 85,502 unique patients remained for the evaluation against cancer patients from the QOR.

The contingency table showing the distribution of matches and mismatches between Medtex and the QOR at a patient-level is shown in Table 1.

**Table 1.** Contingency table of patient matches and mismatches between Medtex and the Queensland Oncology Repository (QOR).

<table>
<thead>
<tr>
<th>QOR</th>
<th>Medtex Patient Notifiable</th>
<th>Medtex Not Notifiable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Patient Exists</td>
<td>12,799</td>
<td>11,021</td>
<td>23,820</td>
</tr>
<tr>
<td>Cancer Patient Not Exist</td>
<td>3,178</td>
<td>58,504</td>
<td>61,682</td>
</tr>
<tr>
<td>Total</td>
<td>15,977</td>
<td>69,525</td>
<td>85,502</td>
</tr>
</tbody>
</table>

As there were a high number of discrepancies between the information contained in the QOR and the classification output by Medtex, a stratified sample of 1,000 patients was randomly selected from the quadrants of the contingency table for manual review by an experienced cancer clinical coder. The breakdown of the patients and their reports are summarised in Table 2.

**Table 2.** Breakdown of the number of patients and reports for manual coding.

<table>
<thead>
<tr>
<th>QOR Patient Exist / Medtex Notifiable</th>
<th># Patients</th>
<th># Reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>QOR Patient Not Exist / Medtex Notifiable</td>
<td>350</td>
<td>638</td>
</tr>
<tr>
<td>QOR Patient Exist / Medtex Not Notifiable</td>
<td>350</td>
<td>469</td>
</tr>
<tr>
<td>QOR Patient Not Exist / Medtex Not Notifiable</td>
<td>150</td>
<td>180</td>
</tr>
<tr>
<td>Total</td>
<td>1,000</td>
<td>1,609</td>
</tr>
</tbody>
</table>

A summary of the results and analyses are reported in the following section. The aim of the analyses is two-fold: 1) to generate an adjusted contingency table to determine the indicative performance of the system, and 2) identify insights and system limitations for ongoing research and development. The classification results from the cancer clinical coder and Medtex and associated detailed error analyses can be made available upon request.
1. Analysis of Medtex patient notifiable classification and patient exists in the QOR

The manual review of 150 patients that exist in the QOR and were also classified as “Cancer Notifiable” by Medtex revealed that there were 8 patients whose reports did not suggest the need for a cancer notification; these patients existed in the QOR but for cancers diagnosed in other years (not in 2009). An error analysis on these Medtex misclassified cases were observed to fall into two categories, namely, 1) Medtex misclassifying SCC/BCC of the skin, which are to be excluded from Cancer Registry notifications, and 2) other algorithm related issues from Medtex such as the missed detection of negative assertion phrases. Due to the 8 false positive cases, there is an error rate of 8/150 = 0.0533 (95% Confidence Interval (CI): 0.025-0.1059). Extrapolating these false positive findings within the respective patient stratum revealed that 682 patients (i.e. error rate x 12,799; 95% CI: 320-1,356) were Medtex notifiable but indeed did not have cancer; thus requiring adjustment.

2. Analysis of Medtex patient not notifiable classification and patient not exist in the QOR

The manual review of 150 patients that did not exist in the QOR and were also classified as “Not Cancer Notifiable” by Medtex revealed that there was 1 patient whose reports did suggest a cancer notification. Upon consultation with the QCR, it was confirmed that the patient case would indeed be not notifiable. As a result, no adjustments were required for patients within this stratum where the patient did not exist in the QOR and Medtex patient is not notifiable.

3. Analysis of Medtex patient not notifiable classification and patient exists in the QOR

The manual review of 350 patients that exist in the QOR but were classified as “Not Cancer Notifiable” by Medtex revealed that there were:

- 9 patients whose reports did suggest the need for a cancer notification,
- 6 patients whose reports contained supplementary-only material for cancer notifications, and
- 335 patients whose reports did not suggest the need for a cancer notification.

An analysis on the 335 patients in the QOR with non-notifiable reports revealed that 30 patients were diagnosed in 2009 but had non-notifiable 2009 reports due to patients being either 1) from the private sector or from another Australian state, 2) having cancer notifications from other sources beyond histology and cytology reports, for example, hospital notifications, and 3) having incomplete records due to the inaccessibility of reports occurring in early 2009 due to pathology feeds only being fully operational from mid-March 2009. The remaining 305 patients in the QOR with non-notifiable 2009 reports were cancers diagnosed in other years. Due to the 335 true negative cases, there is an adjustment rate of 335/350 = 0.9571 (95% CI: 0.9287-0.9749). Extrapolating these true negative findings within the respective patient stratum revealed that 10,549 patients (i.e., adjustment rate x 11,021; 95% CI: 10,235-10,744) were correctly classified as Medtex not notifiable; thus requiring adjustment.

With regards to the 9 patients where Medtex did misclassify reports as “Not Cancer Notifiable”, these errors were observed to broadly fall into 3 categories, namely, 1) Free text to SNOMED CT mapping errors, 2) Incorrect implementation of Cancer Registry business rules, and 3) other algorithm related issues from Medtex such as the incorrect application of negative assertion phrases to SNOMED CT concepts. These algorithmic issues can be the target of future refinements and therefore may be possible to rectify and resolve through the updating of business rules and improvements to the Medtex free text to SNOMED CT mapping engines and negation detection algorithms. Due to the 9 false negative cases, there is an error rate of 9/350 = 0.0257 (95% CI: 0.0126-0.05). Extrapolating these false negative findings within the respective patient stratum revealed that 283 patients (i.e., error rate x 11,021; 95% CI: 139 to 551) were Medtex not notifiable but do contain cancer. As these cases are within the correct patient stratum where Medtex classified a patient incorrectly as “Not Cancer Notifiable”, no further adjustments were required for these errors. These cases may be considered as high importance as these represent missed cancer cases. But as mentioned above, these errors can be potentially rectified or resolved through further system improvements.

4. Analysis of Medtex patient notifiable classification and patient not exist in the QOR

The manual review of 350 patients that do not exist in the QOR but were classified as “Cancer Notifiable” by Medtex revealed that there were:

---

1 95% confidence interval calculated using Wilson’s procedure (including continuity correction; http://www.vassarstats.net/prop1.html)
• 82 patients whose reports did suggest the need for a cancer notification,
• 23 patients whose reports only contained supplementary-only material for cancer notifications (as well as another 68 classifications where both Medtex and the cancer clinical coder assigned patients as “Supporting Notifiable Report”), and
• 125 patients whose reports did not suggest the need for a cancer notification.

An error analysis on the 125 patients that Medtex misclassified as “Cancer Notifiable” revealed that the majority of cases were misclassified into particular tumour streams such as skin (misclassifying melanoma and SCC/BCC of skin; Medtex was unable to ascertain whether the SCC/BCC was associated with “skin”), brain and blood cancers. Due to the 125 false positive cases, there is an error rate of 125/350 = 0.3571 (95% CI: 0.3073-0.4101). Extrapolating these false positive findings within the respective patient stratum revealed that 1,135 patients (i.e., error rate x 3,178; 95% CI: 977 to 1303) were Medtex notifiable but do not contain cancer. As these cases are within the correct patient stratum where Medtex classified a patient incorrectly as “Cancer Notifiable”, no further adjustments are required for these errors.

Regarding the 82 patients who were not in the QOR but were indeed cancer notifiable patients, 9 of the patients were recorded in the QOR with different names and therefore could not be computationally matched. Of the remaining 73 patients who could not be found in the QOR, 52 cases related to leukemia, prostate and cervix cancers. Others were distributed with low frequency counts across other tumour streams. In terms of histological type, 57 out of the 73 patients had cancers in the following four morphologies: SCC, CIN 3, adenocarcinoma and chronic lymphocytic leukaemia, B-cell type. Others were distributed with low frequency counts across other morphologies. Due to the 82 true positive cases, there is an adjustment rate of 82/350 = 0.2343 (95% CI: 0.1916-0.2829). Extrapolating these true positive findings within the respective patient stratum revealed that 745 patients (i.e., adjustment rate x 3,178; 95% CI: 609 to 899) were Medtex notifiable and patients were in fact “Cancer Notifiable”. Note that 663 (i.e., 73/350 x 3,178; 95% CI: 534 to 813) out of the 745 (95% CI: 609 to 899) “Cancer Notifiable” patients were not found in the QOR. These cases may be considered as high importance as these cases result in unreported cancers.

**Adjusted contingency table and overall system performance**

In light of the analysis of the comparisons between the QOR, Medtex, and an expert clinical coder, adjustments to Table 1 were required in order to reflect the true performance of Medtex. In particular, the following adjustments were required:

• 682 (95% CI: 320 to 1356) “Cancer Notifiable” patients identified by Medtex but were “Not Cancer Notifiable”
• 10,549 (95% CI: 10,235-10,744) “Not Cancer Notifiable” patients identified by Medtex who are in the QOR but did not have cancer notifiable 2009 pathology reports, and
• 745 (95% CI: 609 to 899) “Cancer Notifiable” patients identified by Medtex who were not in the QOR but indeed had cancers.

Furthermore, to take into account of “Supporting Notifiable Report” classifications, which may over estimate the number of “Cancer Notifiable” cases during the patient-level consolidation process by Medtex, this effect was accounted for by further adjusting the contingency table, especially in the patient stratum where Medtex classified a patient as “Cancer Notifiable” but the patient did not exist in the QOR; here there were 68 out of the 350 cases where both Medtex and the cancer clinical coder assigned patients as “Supporting Notifiable Report” within the respective patient stratum. As a result, there is an adjustment rate of 68/350 = 0.1943 (95% CI: 0.155-0.2405). Extrapolating this result within the respective patient stratum reveals that 617 patients (i.e., adjustment rate x 3,178; 95% CI: 493 to 764) were Medtex “Cancer Notifiable” but were technically “Supporting Notifiable Reports”. A further adjustment is therefore applied to the contingency table:

• 617 (95% CI: 493 to 764) Medtex “Cancer Notifiable” patients who were not in the QOR but are technically “Not Cancer Notifiable” due to “Supporting Notifiable Report” classifications, indicating either suspected or re-excisions with no residual cancers.

Table 3 presents the adjusted contingency table and Table 4 presents a summary of Medtex’ performance in terms of sensitivity, specificity, positive predicted value and F1 score.
Table 3. Adjusted contingency table of results comparing Medtex and the QOR based on analysis of discrepancies from a manual review.

<table>
<thead>
<tr>
<th></th>
<th>Medtex</th>
<th>Patient Notifiable</th>
<th>Patient Not Notifiable</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual QOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Notifiable</td>
<td>12799 (-682, +745)</td>
<td>= 12,862 (TP)</td>
<td>[Min: 12,052; Max: 13,378]</td>
<td>13,334</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11021 (-10549)</td>
<td>= 472 (FN)</td>
<td></td>
</tr>
<tr>
<td>Patient Not Notifiable</td>
<td>3178 (+682, -745, -617)</td>
<td>= 2,498 (FP)</td>
<td>[Min: 1,835; Max: 3,432]</td>
<td>72,168</td>
</tr>
<tr>
<td></td>
<td></td>
<td>58504 (+10549 +617)</td>
<td>= 69,670 (TN)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15,360</td>
<td>70,142</td>
<td>85,502</td>
</tr>
</tbody>
</table>

*Numbers reported for QOR are “virtual” numbers extrapolated from an analysis of discrepancies between Medtex and the QOR on 2009 pathology data. Min and Max are computed from the lowest and highest combinations of 95% confidence interval (CI) uncertainties, respectively.

Table 4. Medtex’ performance on classifying cancer notifiable patients.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>96.5%</td>
<td>96.5%</td>
<td>83.7%</td>
<td>89.6%</td>
</tr>
<tr>
<td></td>
<td>(94.5-97.8%)</td>
<td>(95.3-97.4%)</td>
<td>(79.6-86.8%)</td>
<td>(86.4-91.9%)</td>
</tr>
</tbody>
</table>

95% confidence interval (CI) uncertainties reported in brackets.

Discussion

Medtex achieves high sensitivity and specificity in identifying cancer notifiable patients from histology and cytology HL7 reports in line with the results from previous work using a smaller and unrepresentative dataset. As a consequence of Medtex’ high sensitivity, the system was able to identify cancer patients who were not found in the QOR. The high sensitivity, however, was at the expense of PPV. Although the effect of “Supporting Notifiable Report” classifications on false positives was partially taken into account, the lower PPV is considered to be of lower priority (or importance) as these cases can be quickly reviewed manually. The observed lower PPV finding was only possible through this larger scale study with a larger range of report types, tumour streams, pathology laboratories and pathologists; previous work on a smaller dataset did not reveal this. Error analysis reveals that system errors (both false negative and false positives) were mainly tumour stream dependent and therefore could be targeted as part of future work. Future work would also include the evaluation of cancer notifications at a cancer case level to ensure that “Cancer Notifiable” patients were correctly matched to their specific cancers.

Unlike other approaches to cancer notifiable classification that rely on custom phrases, explicitly mentioned disease codes, or the development of tumour specific classification models, Medtex aims to be generic and relies on Cancer Registry business rules and the normalisation and reasoning of records using standard clinical terminologies, specifically SNOMED CT. Although, the symbolic rule-based approach used in Medtex was developed from a small development data set, the current study shows how the utility of such a system can be assessed when applied on larger datasets. The findings from this larger dataset study can be used to bootstrap the development of the symbolic rule-based system. Machine-learning approaches to classification require larger training datasets to generalise but labelled data can be difficult to obtain especially from busy and highly skilled health personnel.

The Medtex processing of reports can also be applied at pathology laboratories to assist with fulfilling mandatory cancer notification requirements. Medtex could be used by pathology laboratories to accurately and automatically flag cancer reports for notification to cancer registries. Not only would this benefit cancer registries, but also individual pathology laboratories; thus improving cancer notification management.

Medtex also has the added advantage of producing uniform data, reducing manual processing and assisting in preventing inaccuracies in the data. This can allow cancer registries and pathology laboratories to devote staffing resources to other high value adding areas such as quality assurance, data integrity and manage the registry in a cost effective manner.
Conclusion

The utility of a medical text analytic system, Medtex, on automating Cancer Registry notifications from narrative histology and cytology HL7 reports from an Australian state-wide pathology repository was assessed. The system was developed using Cancer Registry business rules, natural language processing and symbolic reasoning using the SNOMED CT ontology. Cancer report notifications were consolidated at a patient level and compared against patients with notifiable cancers from the Queensland Oncology Repository (QOR).

A stratified random sample of 1,000 patients was manually reviewed by an expert clinical coder to analyse agreements and discrepancies. Upon extrapolation of the results to the full dataset, sensitivity of 96.5% (95% CI: 94.5-97.8%), specificity of 96.5% (95.3-97.4%), positive predictive value of 83.7% (79.6-86.8%), and an F1 score of 89.6% (86.4-91.9%) were achieved for identifying cancer notifiable patients. Medtex was also able to identify cancer patients that were not found in the QOR; resulting in unreported cancers. Error analysis revealed that system errors tended to be tumour stream dependent with specific cancer sites and histological types causing a large portion of the errors. The evaluation of Medtex on large-scale medical records enabled insights for assessing the utility of the medical text analytic system as well as highlighting avenues for future system developments.

The system is currently being piloted within the Department of Health, Queensland Government, to process the historical and new incoming electronic HL7 pathology reports, not only for identifying cancer notifiable cases, but also the automatic extraction of cancer notifications data including cancer stage.

The Medtex processing of other sources of cancer notification including feeds from private pathology laboratories, radiology reports and death certificates is also currently being investigated. Medtex can also be applied to other medical domains beyond cancer.

Acknowledgements

This research was done in partnership between the Australian e-Health Research Centre (AEHRC) within CSIRO and the Queensland Cancer Control Analysis Team (QCCAT) within the Department of Health, Queensland Government. The authors acknowledge Stephen Armstrong and Nilesh Mendis from QCCAT for data access as well as help in getting Medtex deployed on the Department of Health’s infrastructure.

References


Tracking the Remodeling of SNOMED CT’s Bacterial Infectious Diseases

Christopher Ochs, PhD¹, James T. Case, DVM, PhD², Yehoshua Perl, PhD¹
¹New Jersey Institute of Technology, Newark, NJ; ²NLM/NIH, Bethesda, MD

Abstract
SNOMED CT’s content undergoes many changes from one release to the next. Over the last year SNOMED CT’s Bacterial infectious disease subhierarchy has undergone significant editing to bring consistent modeling to its concepts. In this paper we analyze the stated and inferred structural modifications that affected the Bacterial infectious disease subhierarchy between the Jan 2015 and Jan 2016 SNOMED CT releases using a two-phased approach. First, we introduce a methodology for creating a human readable list of changes. Next, we utilize partial-area taxonomies, which are compact summaries of SNOMED CT’s content and structure, to identify the “big picture” changes that occurred in the subhierarchy. We illustrate how partial-area taxonomies can be used to help identify groups of concepts that were affected by these editing operations and the nature of these changes. Modeling issues identified using our two-phase methodology are discussed.

Introduction
SNOMED CT [1] is a large and complex medical terminology that is maintained by the International Health Terminology Standards Development Organization (IHTSDO) [2]. Extensive editorial work goes into each SNOMED CT release. Additional content, in the form of new concepts and new relationships, is added and existing content is revised to address errors and inconsistencies, while content that is no longer valid is retired.

The IHTSDO sometimes focuses editorial resources on a specific portion of SNOMED CT’s content. The Bacterial infectious disease subhierarchy, with over 2000 concepts, underwent significant remodeling in the July 2015 and Jan 2016 releases. The editorial staff at the IHTSDO determined that the concepts in this subhierarchy were modeled inconsistently. The goal of the remodeling is to make stated changes that enable SNOMED CT’s classifier to better infer correct hierarchical relationships. The still ongoing revision has affected hundreds of the Bacterial infectious disease concepts, requiring thousands of individual editing operations. SNOMED CT’s editors and end users need to be aware of the overall impact of these modifications on the content of the terminology. Editors need to ensure that all of their modifications were correct and end users need to know that their systems will still work correctly.

In this paper we analyze how the current, but incomplete modifications to the concepts in the Bacterial infectious disease affected the subhierarchy. However, the task of analyzing changes across multiple SNOMED CT releases is not straightforward. The IHTSDO does not provide a “user friendly” change log. A list of individual changes (a “delta”) is provided with each release, but one editing operation may result in multiple entries. The delta provides no additional information about the change (e.g., is the new target of a relationship more specific, less specific, or unrelated to the old target?), or what motivated it. Additionally, a list of changes, as provided by the delta, does not provide the “big picture” of what changed. When hundreds of concepts are modified it is difficult to review each individual change. As a result, an editor or end user may decide to avoid an in-depth review of the changes. However, such a review is critical, not just to ensure the correctness of each change, but also because some changes result in unintended consequences, causing new errors and inconsistencies, as demonstrated in Wang et al. [3].

To support the identification and analysis of structural changes between SNOMED CT releases we introduce a two-phase summarization-based approach. First, we introduce a method to create a human-readable SNOMED CT change log (based on the delta files provided with each release) that captures and summarizes the individual structural changes that occurred in the stated and inferred versions of SNOMED CT. Second, we utilize partial-area taxonomies[4], which summarize structurally and semantically similar concepts, to capture a summary of the major changes that occurred to groups of these similar concepts. We use this two-phase approach to identify significant structural changes in the Bacterial infectious disease subhierarchy. Between the July 2015 and Jan 2016 releases we identified thousands of editing operations applied to hundreds of concepts. We illustrate how the remodeling significantly affected, and continues to affect, the content of the subhierarchy.

Background

SNOMED CT
The Jan 2016 release of SNOMED CT consists of over 300,000 active concepts and over 1.5 million relationships among these concepts. SNOMED CT is released as a series of tab-delimited files that represent the concepts,
relationships, and descriptions of the terminology. SNOMED CT is based on Description Logics and it is distributed with both the stated relationships, which were explicitly defined by SNOMED CT editors, and inferred relationships, which were obtained by applying a reasoner to the stated relationships. Each SNOMED CT release comes with a “snapshot” of the terminology, which contains all the content for a given release, and with a “delta,” which identifies the individual changes that occurred between the last two releases. A full history of changes for every concept is also provided.

Various studies have looked at SNOMED CT’s content. Ceuster’s [5] analyzed changes in 18 releases of SNOMED CT and identified the need for better documentation of changes. Specifically, documentation on what motivated SNOMED CT authors to introduce changes of a certain type was identified as important. Rector et al. [6] and Mortensen et al. [7] looked at quality issues in SNOMED CT content.

**Partial-area Taxonomies**

We define an Abstraction Network [8] as a hierarchy of nodes, where each node represents a set of structurally similar concepts. It serves as a compact summary of a terminology’s content and structure. In previous work [4, 9-11], we have developed different Abstraction Networks to summarize biomedical terminologies and ontologies, such as the National Cancer Institute thesaurus (NCIt) [12], the Gene Ontology (GO) [13], and SNOMED CT. We have shown that Abstraction Networks support various SNOMED CT use cases, such as quality assurance (QA) [4], characterizing the modeling complexity of SNOMED CT [14], and observing changes to content due to QA [15].

One Abstraction Network we developed to support SNOMED CT QA is the partial-area taxonomy [4, 11], which summarizes groups of structurally and semantically similar concepts. We define a subject subtaxonomy as a partial-area taxonomy rooted at concept that represents a subject area (e.g., Bacterial infectious disease). In Ochs et al. [4] we looked at the properties of subject subtaxonomies for ten important Clinical finding subhierarchies (e.g., Cancer and Heart disease). We will now illustrate the derivation of a subject subtaxonomy using an excerpt from the Jan 2016 release of SNOMED CT using Bacterial infectious disease as the selected subject concept.
Given a SNOMED CT subhierarchy rooted at a concept $c$, we define an area as the set of all concepts in the subhierarchy with the same set of attribute relationship types. Only the types of the relationships are considered; target concepts are disregarded. Areas are named after their set of attribute relationship types (e.g., \{Causative agent, Pathological process\}). We say that the set of concepts summarized by an area are structurally similar, as the concepts are modeled using the same set of attribute relationship types. All areas are disjoint in terms of the concepts they summarize. We define an area taxonomy as an Abstraction Network composed of areas that are connected by hierarchy child-of links based on the underlying IS-A relationships. We define a root concept of an area as a concept that has no parent concepts in its area. An area $A$ is child-of another area $B$ if a root concept in $A$ has a parent concept in $B$. Figure 1 (b) shows the area taxonomy derived from the concepts in Figure 1 (a).

A partial-area consists of a root concept and all of its descendant concepts in its area. Partial-areas are named after their root and labeled with the total number of concepts summarized (this will be written using parenthesis, e.g., Bacterial sepsis (2)). Partial-areas summarize subhierarchies of semantically similar concepts within each area, since all of the concepts in the partial-area are descendants of the root. Multiple partial-areas may summarize the same concept. A partial-area taxonomy is an Abstraction Network, consisting of partial-area nodes, that refines an area taxonomy by explicitly identifying these subhierarchies of semantically similar concepts. Figure 1 (c) shows the partial-area taxonomy (subject subtaxonomy, since it is rooted at a chosen subject concept) for the concepts in Figure 1 (a). We note that partial-area taxonomies can be derived using inferred or stated relationships. To derive partial-area taxonomies we developed a software tool called BLUSNO [16], which provides a user interface for exploring partial-area taxonomies across multiple SNOMED CT releases.

**Terminology Diffs**

We note that there has been extensive work in the area of detecting changes between terminology releases (“diffs,” e.g., [17, 18]). As noted earlier, SNOMED CT provides pre-computed diff information with each release in the form of delta files. In [19] we observe that terminology diff tools often output an overwhelming amount of information (i.e., tens of thousands of identified changes) when many concepts are modified. The two-phase approach described in this paper visually identifies changes to sets of similar concepts, reducing the amount of information one has to view, and provides a more descriptive list of changes than is provided by many diff tool.

**Methods**

To determine how SNOMED CT’s content changed it is necessary to have a list of the editing operations that were applied to the concepts. SNOMED CT’s editors do not maintain such a list and there is no record of justifications for changes. We look at the stated changes SNOMED CT’s editors applied and their effect on the inferred version of the terminology. The delta files provided with each release are not “human readable” and one editing operation may be represented by multiple entries. For example, in the stated version of the July 2015 release, the concept Bacillus infection due to Bacillus had a parent Disease due to Gram-positive bacteria that was replaced by a more general parent, Disease. In the delta file this is expressed as two changes: the removal of the IS-A relationship to Disease due to Gram-positive bacteria and the addition of the IS-A relationship to Disease. There are thousands of entries in the delta file due to the changes in the Bacterial infectious disease subhierarchy. While a list of changes will provide a glimpse at the structural editing operations that were applied, such a list does not give the “big picture” of the overall impact of the changes. Thus, a methodology for detecting and visualizing the overall impact of the changes in the inferred version of SNOMED CT is also necessary.

**Editing Operation Detection and Summarization**

A relationship $r_i$ we abbreviate this as $(s_i, t_i, d_i, g_i)$. In the following we refer to a relationship that no longer exists in a release as an inactive relationship and we refer to a new relationship in a release as an activated relationship. Given the set of activated and inactive relationships, as expressed in one release’s delta files, we can infer the editing operations that were applied and create a “human readable” log of changes that also reduces the overall complexity of the change information. Table 1 identifies five kinds of editing operations that can be detected in the delta file.

We distinguish between hierarchical changes and changes to attribute relationships. We separate editing operations 1–4 in Table 1 into operations applied to IS-A relationships and operations applied to attribute relationships (e.g., added parent and added attribute relationship). These editing operations can be computed using the delta for the stated relationships (i.e., editing operations applied by a SNOMED CT editor) or from the delta for the inferred relationships (i.e., the implicit changes that were inferred by the classifier).
We note that it is possible to identify other kinds of editing operations using the approach described in Table 1. For example, Tuberculous dactylitis has a removed Finding site relationship to Finger structure and an added Finding site relationship to Entire digit. This could be considered a “changed attribute relationship target” operation. In this study we found relatively few removed relationship editing operations. Thus, for this study, we do not explicitly identify this kind of operation. From the delta files we can also infer why some changes occurred. For example, if a destination concept in a relationship is deactivated then it would necessitate that the target change, if the kind of relationship is still needed to model the source concept.

**Table 1.** Five kinds of editing operations which can be identified from SNOMED CT’s relationship delta file

<table>
<thead>
<tr>
<th>Editing Operation</th>
<th>Description</th>
<th>Example (from stated relationships delta file)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Added Relationship</td>
<td>Given an activated relationship ( r_1 ) there exists no inactive relationship ( r_2 ) where ( s_1 = s_2, t_1 = t_2 ) and ( d_1 = d_2, d_1 ) is also not an ancestor or descendant of ( d_2 ).</td>
<td>Carbuncle of ear had a Causative agent relationship added with a target Superkingdom Bacteria.</td>
</tr>
<tr>
<td>2 Removed Relationship</td>
<td>Given an inactive relationship ( r_1 ) there exists no activated relationship ( r_2 ) where ( s_1 = s_2, t_1 = t_2 ) and ( d_1 = d_2, d_1 ) is also not an ancestor or descendant of ( d_2 ).</td>
<td>The Finding site relationship to Finger structure was removed from Tuberculous dactylitis and replaced with “Entire digit”.</td>
</tr>
<tr>
<td>3 More Refined Target</td>
<td>Given an inactive relationship ( r_1 ) there exists an activated relationship ( r_2 ) where ( s_1 = s_2, t_1 = t_2 ) and ( d_2 ) is a descendant of ( d_1 ).</td>
<td>The target of Staphylococcal enterocolitis’s Finding site relationship changed from Intestinal structure to Colon structure and Small intestinal structure.</td>
</tr>
<tr>
<td>4 Less Refined Target</td>
<td>Given an inactive relationship ( r_1 ) there exists an activated relationship ( r_2 ) where ( s_1 = s_2, t_1 = t_2 ) and ( d_2 ) is an ancestor of ( d_1 ).</td>
<td>Bacterial infection due to Bacillus’s stated parent changed from Disease due to Gram-positive bacteria to Disease to Disease.</td>
</tr>
<tr>
<td>5 Relationship Group Changed</td>
<td>Given an activated relationship ( r_1 ) there exists an inactive relationship ( r_2 ) where ( s_1 = s_2, t_1 = t_2 ) and ( d_1 = d_2 ) but ( g_1 ) does not equal to ( g_2 ).</td>
<td>The stated Causative agent relationship from Plague meningitis to Yersinia pestis was moved from group 1 to group 2.</td>
</tr>
</tbody>
</table>

Below is an example of converting two stated relationship delta entries (left) into a summarized editing operation.

<table>
<thead>
<tr>
<th>Active</th>
<th>Source</th>
<th>Type</th>
<th>Destination</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>416265003</td>
<td>116680003</td>
<td>15202009</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>416265003</td>
<td>116680003</td>
<td>64572001</td>
<td>0</td>
</tr>
</tbody>
</table>

**Tuberculoma of brain**
Less refined parent

Tuberculoma => Disease

To enable our editing operation analysis we created a module for the BLUSNO system that is able to process the delta files and output the editing operations. Output is provided as tab delimited text.

**Reflecting on Change Using Subject Subtaxonomies**

Subject subtaxonomies summarize groups of structurally and semantically similar concepts, all of which are related to a chosen subject concept (e.g., Bacterial infectious diseases). By comparing the areas and partial-areas in the partial-area subtaxonomies derived for multiple SNOMED CT releases, one obtains a summary of the major changes that occurred. Consider the examples in Figure 2, which capture how many of the 372 concepts in the Jan 2015 Bacterial infection by site (372) partial-area, in the \{Causative agent, Finding site, Pathological process\} area, changed in July 2015 and Jan 2016. We identify several significant changes based on the subject subtaxonomies in Figure 2. These changes result in modifications to the areas and partial-areas in the subject subtaxonomy.

**Resulting introduced partial-areas.** One major change in July 2015 was the introduction of 110 partial-areas to the \{Causative agent, Finding site, Pathological process\} area (19 are shown in Figure 2(b)). The concepts in these 110 partial-areas were summarized by Bacterial infection by site (372) in Jan 2015. These concepts (e.g., the root of the Tuberculosis of gastrointestinal tract (9) partial-area in Figure 2(b)) did not change structurally (i.e., they have same set of attribute relationships in both releases) but they did change semantically (i.e., the hierarchy changed and they are no longer descendants of Bacterial infection by site).
Figure 2. An example of how the changes to the concepts in the Bacterial infection by site (372) partial-area are captured in subject subtaxonomies derived for three SNOMED CT versions. (a) The Jan 2015 ‘Causative agent, Finding site, Pathological process’ area with its single partial-area. (b) The July 2015 version of the area from (a). The partial-area Bacterial infection by site lost 196 concepts and a total of 110 new partial-areas (19 shown) appeared due to changes in the underlying concept hierarchy. (c) Jan 2016 version of the area from (a). The blue partial-areas in (b) were merged back into the Bacterial infection by site partial-area. New partial-areas (e.g., Tetanus (9)) also appeared, indicating a change in their concept’s semantics. Several partial-areas (e.g., Lepromatous leprosy (2)) still exist in the area. Other partial-areas moved to areas with more relationships (e.g., Tuberculosis of genitourinary system (6) moved to the area ‘Associated morphology, Causative agent, Finding site, Pathological process’) and grew to 19 concepts). The concepts in Tuberculosis of urinary organs (4) moved to this partial-area.

There were no stated editing operations applied to Tuberculosis of gastrointestinal tract. Thus, one needs to look at the inferred changes. Three such changes occurred: the classifier inferred a less refined parent (Disorder of gastrointestinal tract), a new Causative agent attribute relationship to Genus Mycobacterium, and a Pathological process attribute relationship was moved to a relationship group with the inherited Causative agent relationship. The stated editing operations that lead to this change were applied to Mycobacteriosis, the parent of Tuberculosis.

By investigating the root concepts of the other new partial-areas in Figure 2(b) we observe a similar phenomenon. All of the (former) descendant concepts of Bacterial infection by site that did not have the same relationship group organization (i.e., the Finding site attribute was not assigned to a relationship group with Causative agent and Pathological process attribute relationships), were no longer classified under Bacterial infection by site after July 2015, and thus, appeared as new partial-areas. Looking at the Jan 2016 version of this area in Figure 2(c), one observes that some of these partial-areas are still in the area (e.g., Lepromatous leprosy (2)). One can also see that several partial-areas were added (e.g., Tetanus (9)), again indicating that sets of concepts are no longer being inferred as descendants of Bacterial infection by site.

**Resulting removed partial-areas.** Some of the partial-areas introduced in the July 2015 ‘Causative agent, Finding site, Pathological process’ area no longer exist in the area in Jan 2016. The concepts in certain partial-areas (colored blue in Figure 2(b)) returned to Bacterial infection by site in Jan 2016 due to some combination of editing operations. Other partial-areas, however, moved to entirely different areas, indicating a change in structure.

**Resulting changes to areas.** When a concept moves from one area to another it indicates a change in the structure of the concept. Specifically, it means additional (or fewer) types of attribute relationships are used to define the concept. For example, the partial-area Tuberculosis of bones and/or joints (20), in ‘Causative agent, Finding site, Pathological process’ in July 2015, moved to ‘Associated morphology, Causative agent, Finding site, Pathological process’ in Jan 2016. Additionally, eight concepts (e.g., Tuberculosis of hip and Tuberculosis synovitis) were inferred as belonging to this partial-area’s subhierarchy. In July 2015, four of the eight concepts were summarized by only Bacterial arthritis (33) in ‘Associated morphology, Causative agent, Finding site, Pathological process’ (not shown in Figure 2(b)). In Jan 2016 the concepts are summarized by both the Tuberculosis of bones and/or joints (28) and Bacterial arthritis (37) partial-areas. Others concepts (e.g., Tuberculous necrosis of bone) were summarized by a singleton partial-area (i.e., it has only one concept) in ‘Associated morphology, Causative agent, Finding site, Pathological process’ in July 2015 but were absorbed into Tuberculosis of bones and/or joints (28) in
Jan 2016. Additional, the concepts from Tuberculosis of gastrointestinal tract (9) in July 2015 are in Bacterial gastroenteritis (24) in Jan 2016, indicating a change in both structure and semantics.

From these examples one can see that the differences between subject subtaxonomies can be used to identify major changes to the hierarchy (via the addition and removal of partial-areas, indicating significant modifications and disruptions) and structure of the concepts (via sets of concepts moving between One does not need to review individual concepts. There are typically significantly fewer partial-areas than concepts, thus, one can review significantly less information, saving editorial resources and enabling a better orientation to the impact of the remodeling. When a significant change is identified via the subject taxonomy, only then would an editor need to review the individual changes which occurred. With a better orientation into the global impact of the changes one will have a greater chance of discovering unwanted side effects of the remodeling and focus future editorial work on correcting the incorrect inferences.

Principles Guiding the Remodeling of Bacterial Infectious Diseases

The above methodologies, implemented in the BLUSNO software tool, allow a user to identify what changed in a summarized manner. However, to understand why the content changed, one needs to look at and understand some of the modeling principles followed by IHTSDO’s editors. The end goal of the remodeling effort was to enable SNOMED CT’s classifier to infer most of the hierarchical connections between concepts in the subhierarchy. This will save SNOMED CT’s editors significant time and effort. Furthermore, the prior modeling of the Bacterial infectious disease concepts led to incorrect inferences that affected the quality of the content.

No proximal primitive parent: One common design pattern in SNOMED CT is the use of a closest proximal primitive parent in the stated modeling of a concept. A proximal primitive parent is the closest parent in the hierarchy to a concept being modeled that is not fully defined [20]. For example, in the Jan 2016 stated version of SNOMED CT, the closest proximal primitive parent of the concept Furuncle of face is Disease. By modeling concepts using the closest proximal primitive parent design pattern, the SNOMED CT classifier can more accurately infer parents and subtypes. In Jan 2015 few of the concepts in the Bacterial infectious disease hierarchy followed this pattern. The major focus of the hierarchy redesign has been to consistently apply this pattern to the content.

Many of the concepts in the Bacterial infectious disease subhierarchy (1404, 68.9% in Jan 2015) that did not follow the proximal primitive parent pattern had one or more stated parents that were fully defined. These parents are unnecessary and often cause incorrect inferences. When these concepts are modeled with a Pathological process relationship with a target of Infectious process and a Causative agent with a target of some kind of bacteria, the classifier auto-classifies the concepts into the Bacterial infectious disease subhierarchy.

Inconsistent relationship grouping: A concept may have multiple relationships of the same type (e.g., multiple Finding site relationships, like Boil of lower limb in Figure 3(b)). To correctly associate sets of related relationships SNOMED CT includes a mechanism to organize relationships into sets called relationship groups [21]. In the Bacterial infectious disease subhierarchy the attribute relationships for many of the concepts were not consistently organized into relationship groups.

A large portion of the subhierarchy’s concepts (1857, 91.2% of the hierarchy) had at least one of the above issues and were considered to be inconsistent. The remodeling process was carried out by one editor, (JTC). The remodeling process was done manually, requiring the editor to modify each concept to fit the model. Several editing operations may be required to remodel a concept. Figure 3 (a) and (b) illustrate the stated remodeling of Boil of lower limb. This concept required a closest proximal primitive parent (Disease) and several stated attribute relationships. Figure 3 (e) and (f) illustrate the remodeling of Staphylococcal infection of skin. The kind of stated remodeling shown in Figure 3 was applied to nearly 1000 concepts between the Jan 2015 and Jan 2016 releases.

Consider again Tuberculosis of gastrointestinal tract. The parent that was removed in July 2015, Bacterial gastrointestinal infectious disease, underwent remodeling to fit the closest proximal primitive parent concept model and all of its attribute relationships were assigned the same relationship group. However, in July 2015 Tuberculosis of gastrointestinal tract was not yet remodeled to fit the closest proximal primitive parent model.

Thus, its Finding site attribute relationship was not in the same relationship group as its Causative agent and Pathological process attribute relationships. Due to this difference, SNOMED CT’s classifier could not infer Bacterial gastrointestinal infectious disease as a parent of Tuberculosis of gastrointestinal tract in July 2015.
Figure 3. Examples of stated modeling for concepts remodeled between Jan 2015 and Jan 2016. (a-d) Show the modeling of Boil of lower limb and (e-h) show the modeling of Staphylococcal infection of skin. (a) Stated, Jan 2015. (b) Stated, Jan 2016. (c) Inferred, Jan 2015. (d) Inferred, Jan 2016. (e) Stated, Jan 2015. (f) Stated, Jan 2016. (g) Inferred, Jan 2015. (h) Inferred, Jan 2016.

In Jan 2016 Tuberculosis of gastrointestinal tract was remodeled to fit the closest proximal parent model. The classifier inferred a more refined parent Bacterial gastroenteritis, a child of Bacterial gastrointestinal infectious disease. Thus, in Jan 2016, Tuberculosis of gastrointestinal tract was inferred with a more specific classification.

By reviewing the partial-areas that appeared in \{Causative agent, Finding site, Pathological process\} in July 2015 and then “merged back” into the Bacterial infection by site partial-area in Jan 2016, we observe that their root concepts all underwent remodeling to fit the closest proximal primitive parent model. Once this remodeling was complete the concepts summarized by these partial-areas were correctly inferred as being descendants of Bacterial infection by site. As a larger proportion of the hierarchy is remodeled in future releases, we expect that the concepts in the remaining partial-areas (e.g., Lepromatous leprosy) will eventually “merge” back into the Bacterial infection by site partial-area or move to another area, indicating that their remodeling has been completed.

Results

We obtained the Jan 2015, July 2015, and Jan 2016 international releases of SNOMED CT in RF2 format [22]. For each release we used our BLUSNO tool to create a subject subtaxonomy rooted at Bacterial infectious disease using each concept’s inferred relationships. For each subject subtaxonomy we determined the area and partial-area(s) that summarized each concept. For the July 2015 and Jan 2016 releases we identified the stated editing operations and inferred changes that affected the subhierarchy’s active concepts. Table 2 provides metrics for the subhierarchy and subject subtaxonomies for each release and illustrates the desired improvements that were expected from the remodeling effort (e.g., a decrease in primitive concepts and a decrease in the use of stated fully defined parents).
Table 2. *Bacterial infectious disease* subhierarchy and subject subtaxonomy metrics across the three releases

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Concepts</td>
<td>2037</td>
<td>2006</td>
<td>2066</td>
</tr>
<tr>
<td># Primitive concepts</td>
<td>847</td>
<td>685</td>
<td>519</td>
</tr>
<tr>
<td># Concepts with a Fully Defined Stated Parent</td>
<td>1404</td>
<td>893</td>
<td>407</td>
</tr>
<tr>
<td># Concepts with No Relationship Group</td>
<td>1368</td>
<td>949</td>
<td>508</td>
</tr>
<tr>
<td># Concepts that Don’t Follow Closest Proximal Primitive Parent Model (approximately)</td>
<td>1857</td>
<td>1214</td>
<td>622</td>
</tr>
<tr>
<td># Areas</td>
<td>26</td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td># Partial-areas</td>
<td>431</td>
<td>573</td>
<td>514</td>
</tr>
</tbody>
</table>

In the July 2015 release we identified 2218 editing operations that affected 541 (/2006 = 27.0%) active concepts in the subhierarchy. In the Jan 2016 release there were 2215 editing operations that affected 560 (/2066 = 27.1%) concepts. There were 17 concepts that had editing operations applied in both releases. In the inferred version of the subhierarchy the stated editing operations led to many inferred changes. In the July 2015 release there were 7209 implicit changes that affected 1791 (/2066 = 86.7%) concepts. In the Jan 2016 release there were 4543 implicit changes that affected 1016 (/2066 = 49.2%) concepts. A total of 742 concepts were affected in both releases. Table 3 lists the various editing operations that were applied to the subhierarchy, and their frequency for each release.

Table 3. Number of concepts with a stated editing operation or affected inferentially during the reasoning process

<table>
<thead>
<tr>
<th>Stated Editing Operations</th>
<th>Inferred Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Added parent</td>
<td>19</td>
</tr>
<tr>
<td>Added attribute relationship</td>
<td>531</td>
</tr>
<tr>
<td>Less refined parent</td>
<td>462</td>
</tr>
<tr>
<td>Less refined attribute relationship</td>
<td>14</td>
</tr>
<tr>
<td>More refined parent</td>
<td>4</td>
</tr>
<tr>
<td>More refined attribute relationship</td>
<td>9</td>
</tr>
<tr>
<td>Relationship group changed</td>
<td>141</td>
</tr>
<tr>
<td>Removed parent</td>
<td>98</td>
</tr>
<tr>
<td>Removed attribute relationship</td>
<td>7</td>
</tr>
</tbody>
</table>

Several editing operations were typically applied to each remodeled concept, with several instances of a given type of editing operation typically applied. For example, *Salmonella infection* had one instance of a “less refined stated parent” editing operation and two instances of an “added attribute relationship” editing operation. Most concepts (346 in July 2015, 365 in Jan 2016) had two kinds of editing operations applied. The most common pair of editing operations was making a stated parent less refined and adding one or more attribute relationships (e.g., the examples in Figure 3). The most common less refined parent was *Disease*, with 494 instances in July 2015 and 533 in Jan 2016. Many concepts had several stated fully defined parents replaced by a single stated *Disease* parent. The second most common stated parent was *Sexually transmitted infectious disease* (a primitive concept).

From Table 2 we observe two important aspects of the changes to the subhierarchy. First, the percentage of primitive concepts decreases significantly between the releases. A higher proportion of fully defined concepts are desirable because it indicates that more concepts are sufficiently defined and are distinguished from their parent concepts. Fully defined concepts can also be used by SNOMED CT’s classifier to infer additional relationships.

Second, we observed that the size of the *Bacterial infectious disease* hierarchy changed between each release. In the July 2015 release 136 concepts were removed from the subhierarchy and 105 concepts were added. Some of this change was the result of deactivating concepts. However, many of the removed concepts (97 total) were still active in July 2015 but were no longer inferred as belonging to the subhierarchy (e.g., *Bacterial tonsillitis*, *Brodie's abscess*, and *Traveler's diarrhea*). The continued remodeling for the Jan 2016 release saw only 34 concepts removed from the subhierarchy and the addition of 94 concepts. Out of the 94 added concepts, 63 were removed in the July 2015 release and are now correctly inferred as belonging to the subhierarchy (e.g., *Traveler's diarrhea*). A total of 71 concepts that were in the subhierarchy in Jan 2015, and are still active in Jan 2016, have not returned to the *Bacterial infectious disease* subhierarchy (e.g., *Brodie's abscess* and *Tuberculosis of abdomen*).
Discussion

The Bacterial infectious disease subhierarchy has undergone significant remodeling over the last year. The same effort is planned for the entire Infectious disease subhierarchy, with 6,239 concepts in Jan 2016. A similar effort is ongoing for the Congenital disease subhierarchy. The benefit of the remodeling effort is that the concepts will follow a consistent model, enabling the classifier to infer most of the hierarchical connections between concepts. The remodeling of the Bacterial infectious disease subhierarchy will reduce the amount of future work needed to maintain the subhierarchy in later releases.

The significant amount of change from this remodeling effort is reflected in the Bacterial infectious disease subject subtaxonomies. From Table 2, we observe that the number of partial-areas increased greatly between the Jan 2015 and July 2015 releases, while the number of areas largely stayed the same and the number of concepts slightly decreased. The number of concepts in each area also did not change significantly. This indicates a significant change in semantics. In the July 2015 subtaxonomy we observed the addition of 214 partial-areas (110 of which were in the {Causative agent, Finding site, Pathological process} area) and the removal of 72 partial-areas. A total of 287 concepts were summarized by new partial-areas while still being in the same area and only 17 concepts were summarized by a new partial-area in a different area, indicating that the amount of structural change was minimal.

Jan 2016 saw fewer added partial-areas (76). A total of 74 concepts were in the same area but in new partial-areas (e.g., Tetanus in Figure 2(c)). In Jan 2016 there were significantly more concepts summarized by new partial-areas in different areas (60, e.g., Tuberculosis of bones and/or joints). In comparison to the July 2015 subject subtaxonomy, there were relatively many removed partial-areas (135 total), summarizing 159 concepts. A total of 106 of these concepts returned to their original Jan 2015 partial-areas (e.g., the blue partial-areas in Figure 2(b)). A total of 93 of the partial-areas removed in Jan 2016 only existed in July 2015. The Jan 2016 release thus saw a greater change in attribute relationship structure. For example, 130 concepts moved to the {Associated morphology, Causative agent, Finding site, Pathological process} area, either as part of their partial-area (e.g., Tuberculosis of bones and/or joints) or to a new partial-area (e.g., Tuberculosis of gastrointestinal tract). These concepts were thus modeled with one or more additional kinds of attribute relationships versus their Jan 2015 modeling.

One aspect of the remodeling that is surprising is how disruptive the process has been. As highlighted by the addition and removal of hundreds of partial-areas over the two releases, there has been significant change to these concepts. The concepts in the 110 partial-areas that “emerged” from the Bacterial infection by site partial-area in the July 2015 release were no longer classified as bacterial infections by site. As additional concepts undergo remodeling they will be reclassified into the Bacterial infection by site subhierarchy, however this process is incomplete and it is expected that this hierarchy will be inconsistent for several more SNOMED CT releases.

The editing tool that was used to remodel the content, the IHTSDO Workbench [23] does not support certain collaborative terminology development functionality, such as branching. Typically, when undertaking a significant development project, a terminology will be “branched” and all modifications are made to a separate copy of the terminology. Only after the entire task has been completed will the changes be integrated into the release version of the terminology. In contrast, for the Bacterial infectious disease remodeling project, incomplete remodeling was released in the July 2015 and Jan 2016 releases.

Conclusions

In this paper we described a two-phase methodology to track important changes in a SNOMED CT subhierarchy. We introduced a method of summarizing the editing operations that were applied to a subhierarchy based on the delta files provided in each SNOMED CT release. We also illustrated how partial-area taxonomies highlight the major changes that occurred in the Bacterial infectious disease subhierarchy. We utilized this methodology to analyze the impact of remodeling the Bacterial infectious disease subhierarchy, whose concepts are being modified to fit a standardized concept model. This methodology can help track remodeling changes in other subhierarchies.
Acknowledgement

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01 CA190779. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health. This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

References

Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research

William Ogallo, RPh., MA, MPhil1, Andrew S. Kanter, MD, MPH, FACMI1

1Department of Biomedical Informatics, Columbia University, New York

Abstract

This paper describes a theory derivation process used to develop a conceptual framework for medication therapy management (MTM) research. The MTM service model and chronic care model were selected as parent theories. Review article abstracts targeting medication therapy management in chronic disease care were retrieved from Ovid Medline (2000-2016). Unique concepts in each abstract were extracted using MetaMap and their pairwise co-occurrence determined. The information was used to construct a network graph of concept co-occurrence that was analyzed to identify content for the new conceptual model. 142 abstracts were analyzed. Medication adherence is the most studied drug therapy problem and co-occurred with concepts related to patient-centered interventions targeting self-management. The enhanced model consists of 65 concepts clustered into 14 constructs. The framework requires additional refinement and evaluation to determine its relevance and applicability across a broad audience including underserved settings.

Introduction

Drug therapy problems are undesirable events or circumstances that involve drugs and interfere with the achievement of desired goals of therapy. They can be categorized according to medication indication (e.g., unnecessary treatment, need for additional treatment), effectiveness (e.g., ineffective drug, dosage too low), safety (e.g., adverse drug reactions, dosage too high), and adherence (non-adherence). Drug therapy problems are often deleterious and costly. In the United States, it is estimated that adverse drug events are responsible for an estimated 3.5 million physician office visits, 1 million emergency department visits and 125,000 hospital admissions per year.

Medication therapy management (MTM) is a standard practice for assessing patient drug-related needs, and identifying and resolving drug therapy problems. MTM services are pharmacist-provided, non-dispensing services that aim to optimize drug therapy and improve clinical outcomes of patients. MTM services are the embodiment of the philosophy of pharmaceutical care. This philosophy asserts that it is the responsibility of the pharmacist to meet all drug-related needs of the patient, to be held accountable for those needs and to assist the patient in achieving his or her therapeutic goals through collaboration with other health professionals. MTM promotes medication adherence, safety, and effectiveness and is useful in improving the patient’s medication experience especially in chronic disease.

A conceptual framework is a meaningfully integrated collection of concepts, concept relationships, and assumptions that broadly explains phenomena of interest, expresses assumptions and reflects a philosophical stance. Conceptual frameworks are constructed to predict or explain the relationships, events or behaviors of phenomena. They are essential for strengthening theory-driven research whose aim is to investigate phenomena and support generalizability of findings. There are few examples of substantiated conceptual frameworks for guiding theoretical research in MTM. Consequently, studies within the MTM domain lack shared definitions and are highly heterogeneous with respect to the methods, predictors, and outcomes of interest. Furthermore, currently available conceptual frameworks such as the MTM Service Model in pharmacy practice describe specific dimensions of care delivery rather than specify relationships among concepts that could be scientifically investigated in hypothesis-driven studies. Therefore, it is important to develop new explanatory conceptual frameworks that could be used to strengthen correlational research within the MTM domain.

A common conceptual framework development strategy is theory derivation. In this process, a parent theory is selected and used to guide the development of a new model supported by evidence from current literature. Theory derivation is useful when related concepts could benefit from a structural representation of their inter-relationships, and when
this insight can inform future research. Conventional theory derivation requires the application of thematic analysis where semantic themes are manually derived by coding literature excerpts and collating the codes into themes. While this is arguably the most logical approach for synthesizing information from text, it is often labor- and time-intensive and is subject to investigator bias. Natural language processing is essential for automating knowledge discovery from literature corpora for decision support, guideline development, and medical literature indexing. Text mining using natural language processing has the advantage of enabling fast and efficient analysis of large corpora of documents and may be less prone to investigator bias. Combining this approach with network analysis may provide a systematic way quantitatively analyzing and visualizing salient aspects in biomedical literature.

This article describes the first iteration of a theory derivation process for the development of a conceptual framework to guide medication therapy management research. The model, entitled Enhanced Medication Therapy Management Evaluation Framework, is based on the MTM Service Model in pharmacy practice and the Chronic Care Model that were selected as parent theories. Concepts incorporated into the enhanced model were extracted from abstracts of systematic review and meta-analysis articles using natural language processing and characterized using network theory. This study reports our current findings.

**Methods**

The theory derivation process applied in this study is a four-step process: a) selection of parent theory, b) literature search and review, c) semantic concept analysis, and d) content development. These are described further below.

**Selection of Parent Theory**

We focused on delineating concepts and relationships at the intersection of medication therapy management and chronic disease care. As such, the MTM Service Model in pharmacy practice and the Chronic Care Model were selected as parent frameworks. The MTM service model includes five core elements: medication therapy review, personal medication record, medication-related action plan, intervention and/or referral for drug therapy problems, and documentation and follow-up. These core elements were established through consensus by a group of seven national pharmacy organizations in the United States in 2008. They form a framework for the delivery of MTM services in pharmacy practice. The MTM service model was used to provide an understanding of the goals, core features, process and outcomes associated with the delivery of MTM services. The chronic care model is an organizing framework for improving the management of chronic diseases. It asserts that comprehensive care improvement requires an approach that incorporates six interdependent components: community resources, health system support, self-management support, delivery system design, decision support and clinical information systems. Researchers have revised this model to demonstrate how eHealth tools can assist patients in managing their own chronic illnesses. The chronic care model was used in this study to provide insight into the role of self-management support, community involvement, healthcare delivery system design, decision support and clinical information systems in chronic disease management.

**Literature search and Review**

One author (WO) conducted a literature search to identify systematic review and meta-analysis articles that covered the topics related to medication therapy management in chronic care. For inclusion, articles must have reported results about antecedent factors associated with drug therapy problems or intervention strategies for drug therapy problems in chronic disease care. Studies were identified by electronically searching the Ovid Medline® database. The search strategy involved using medical subject headings (MeSH) and search strings associated with constructs of the MTM service model and the chronic care model frameworks (Table 1). The search was limited to abstracts of original or review articles published in English between 2000 and 2016. A total of 285 abstracts were identified out of which 12 duplicates were removed. An additional 131 articles were rejected through title and abstract review leaving 142 articles that were included in the analysis corpus.

**Semantic Concept Analysis**

The goal of the semantic concept analysis step was to identify and select the key concepts that could be incorporated into the enhanced conceptual framework. This step involved using natural language processing to extract the concepts reported in the abstracts selected for analysis, computing the frequencies of the identified concepts and concept-concept pairs, and using graph analysis to describe and visualize the resulting information. This analysis was based on the assumption that each biomedical literature abstract consists of a set of concepts that reflect the core contents of
Table 1. Terms and strategy for literature search in Ovid Medline.

<table>
<thead>
<tr>
<th></th>
<th>Searches</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(&quot;Medication Therapy Management&quot; or &quot;Medication Reconciliation&quot;).sh.</td>
<td>1374</td>
</tr>
<tr>
<td>2</td>
<td>(&quot;MTM&quot; or &quot;DTM&quot; or &quot;drug therapy management&quot; or &quot;drug therapy problems&quot; or &quot;medication action plan&quot; or &quot;medication reconciliation&quot; or &quot;medication review&quot; or &quot;medication therapy management&quot; or &quot;medication therapy review&quot; or &quot;medication-related problem&quot; or &quot;medication-related problems&quot; or &quot;personal medication record&quot; or &quot;pharmaceutical care&quot; or ((drug or medication) and adherence) or &quot;adverse drug reaction&quot; or &quot;adverse reaction&quot; or &quot;adverse drug event&quot; or &quot;adverse reaction&quot; or &quot;adverse event&quot;).ti.ab.</td>
<td>39199</td>
</tr>
<tr>
<td>3</td>
<td>(&quot;Self Efficacy&quot; or &quot;Patient Participation&quot; or &quot;Self Care&quot; or &quot;Self-Help Groups&quot; or &quot;Social Support&quot; or &quot;Patient-Centered Care&quot; or &quot;Clinical Decision Support&quot; or &quot;Medical Informatics&quot; or &quot;Electronic Health Records&quot; or &quot;Mobile Applications&quot; or &quot;Cell Phones&quot; or &quot;Telemedicine&quot; or &quot;Text Messaging&quot;).sh.</td>
<td>160439</td>
</tr>
<tr>
<td>4</td>
<td>(&quot;Chronic Care Model&quot; or &quot;CCM&quot; or &quot;community health worker&quot; or &quot;frontline health worker&quot; or &quot; frontline worker&quot; or &quot;lay health worker&quot; or &quot;lay worker&quot; or &quot;outreach worker&quot; or &quot;allied health worker&quot; or &quot;patient activation&quot; or &quot;patient empowerment&quot; or &quot;patient participation&quot; or &quot;peer coach&quot; or &quot;peer counselling&quot; or &quot;peer counselor&quot; or &quot;peer education&quot; or &quot;peer educator&quot; or &quot;peer group&quot; or &quot;peer health worker&quot; or &quot;peer mentor&quot; or &quot;peer support worker&quot; or &quot;peer support&quot; or &quot;peer supporter&quot; or &quot;peer worker&quot; or &quot;peer-led&quot; or &quot;self-efficacy&quot; or &quot;self-management&quot; or &quot;village health worker&quot; or &quot;volunteer health worker&quot; or peer?:?peer or self?:?care or task?:redistribution or task?:shifting or &quot;android app&quot; or &quot;android phone&quot; or &quot;cell phone&quot; or &quot;cellular phone&quot; or &quot;information communication technology&quot; or &quot;mobile app&quot; or &quot;mobile device&quot; or &quot;mobile technology&quot; or &quot;short message service&quot; or text?:msg or app or apps or e?:health or electronic?:health or &quot;handheld computer&quot; or handheld device or &quot;interactive voice response&quot; or IVR or mHealth or MMS or mobile?:health or &quot;reminder system&quot; or &quot;reminder&quot; or smartphone or SMS or telecare or telehealth or telemedicine or &quot;wireless technology&quot;).ti.ab.</td>
<td>66962</td>
</tr>
<tr>
<td>5</td>
<td>(drug or medication or medicament or treatment or prescription or pharmaceutical or dose or remedy or cure or therapy).af.</td>
<td>6711256</td>
</tr>
<tr>
<td>6</td>
<td>animal/ not (human/ and animal/)</td>
<td>4159388</td>
</tr>
<tr>
<td>7</td>
<td>((1 or 2) and (3 or 4) and 5) not 6</td>
<td>2414</td>
</tr>
<tr>
<td>8</td>
<td>limit 7 to (abstracts and english language and &quot;review articles&quot; and yr=&quot;2000 - 2016&quot;)</td>
<td>285</td>
</tr>
</tbody>
</table>

what was studied and reported. Consequently, the frequencies of concepts and co-occurrent concept pairs in a corpus of abstracts about a given topic provide a reasonable quantitative description of the relevant themes related to that topic. Concepts in the abstracts selected for the study were extracted using MetaMap and analyzed using graph theory. MetaMap is a natural language processing program that maps biomedical text to the UMLS Metathesaurus 16. Graph theory is the study of graphs – data structures that represent pairwise relationships between discrete objects 17.

Each of the 142 abstracts selected for the study was preprocessing by removing all non-ASCII characters to enable processing via MetaMap. Next, the MetaMap 2016 Web API was used to process the abstracts. The MetaMap output files were further analyzed in R Statistical Programming Language 18. All identified concepts were reviewed to eliminate those that were not relevant based on their UMLS semantic types. Relevant concepts that had similar or related meanings such as therapy adherence, adherence to medication regime, and medication compliance were identified and merged after looking up concept-concept relationships in a local installation of the UMLS 2015AB Metathesaurus 16. All unique concept-concept pairs in each abstract were identified and their frequencies across the entire corpus computed. The results of this process were a list of all relevant concepts (nodes/vertices), and a list of all unique pairs of co-occurrent concepts (edges) as well as their frequencies (edge weights). These data were used to construct and visualize a weighted graph of concepts and their relationships in Gephi 0.9.1 19. Key concepts were identified by computing the weighted degree centrality and decomposing the graphs at various pairwise co-occurrence frequency cutoffs.
Table 2. Descriptive statistics of the concepts, concept pairs, and graph analyzed in the study.

<table>
<thead>
<tr>
<th>Description</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Concepts</strong></td>
<td></td>
</tr>
<tr>
<td>Total number of text to concept mappings in corpus (n)</td>
<td>27,094</td>
</tr>
<tr>
<td>Number of text to concept mappings per abstract (Median [Range])</td>
<td>172 [56-626]</td>
</tr>
<tr>
<td>Total number of unique concepts in corpus (n)</td>
<td>4,090</td>
</tr>
<tr>
<td>Number of unique concepts per abstract (Median [Range])</td>
<td>149 [56-501]</td>
</tr>
<tr>
<td>Total number of unique concepts in corpus included (n(%)</td>
<td>50(1.2%)</td>
</tr>
<tr>
<td>Number of unique concepts included per abstract (Median [Range])</td>
<td>4 [2-13]</td>
</tr>
<tr>
<td><strong>Concept pairs</strong></td>
<td></td>
</tr>
<tr>
<td>Total number of unique concept pairs in corpus included in graph analysis (n)</td>
<td>102</td>
</tr>
<tr>
<td>Number of unique concept pairs included per abstract (Median [Range])</td>
<td>3 [1-14]</td>
</tr>
<tr>
<td><strong>Graph</strong></td>
<td></td>
</tr>
<tr>
<td>Number of nodes (n)</td>
<td>50</td>
</tr>
<tr>
<td>Number of edges (n)</td>
<td>102</td>
</tr>
<tr>
<td>Weighted Degree Centrality (Mean [Range])</td>
<td>18.64 [1-388]</td>
</tr>
<tr>
<td>Graph Density (%)</td>
<td>8.3%</td>
</tr>
</tbody>
</table>

Table 3. Top 10 concepts realized from graph analysis and ranked by weighted degree centrality (Concept ID is UMLS Unique Concept Identifier or CUI)

<table>
<thead>
<tr>
<th>Concept ID</th>
<th>UMLS Preferred Name</th>
<th>% Occurrence in Abstracts</th>
<th>Weighted Degree Centrality</th>
<th>Degree Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1171369</td>
<td>Adherence</td>
<td>93.7%</td>
<td>388</td>
<td>47</td>
</tr>
<tr>
<td>C0086969</td>
<td>Self-Management</td>
<td>42.6%</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>C2986593</td>
<td>Tailored Intervention</td>
<td>24.6%</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>C0162648</td>
<td>Telemedicine</td>
<td>17.6%</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>C3508152</td>
<td>Patient Engagement</td>
<td>21.8%</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>C0030688</td>
<td>Patient education</td>
<td>19.0%</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>C0546816</td>
<td>Persistence</td>
<td>4.2%</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>C0600564</td>
<td>Self Efficacy</td>
<td>13.4%</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>C0877248</td>
<td>Adverse event</td>
<td>9.9%</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>C0517785</td>
<td>Medication Knowledge</td>
<td>4.2%</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Content development

The final step of the conceptual framework creation was the manual refinement and organization of the concepts identified in the previous steps. This involved the process of concept reduction where abstract constructs were refined to concepts which can be reduced to measurable variables. The Donabedian’s Quality Framework was used to organize the entities of the new conceptual framework meaningfully. The Donabedian model is a model for assessing the quality of health. It is based on the assertion that structures of care influence processes of care, and that the latter influence Health outcomes. Structures of care describe the physical and organizational aspects such as infrastructure, equipment, financing, and personnel. Processes of care are the resources and mechanisms for carrying out healthcare activities. Health outcomes are the goals and effects of healthcare on patients such as recovery, survival, and patient satisfaction.

Results

Semantic Concept Analysis Results

Table 2 provides the summary statistics of the concepts, concept pairs and graph analyzed of the 142 abstracts included in the study. A total of 27,094 text tokens in the corpus were mapped to 4,090 unique UMLS concepts. Of these, only 50 concepts (1.2%) were considered relevant and included. A majority of the excluded concepts resulted from the
Table 4. Top 10 concept pairs ranked by percent proportion of occurrence in the analyzed abstracts (Concept ID is UMLS Unique Concept Identifier or CUI)

<table>
<thead>
<tr>
<th></th>
<th>Concept ID</th>
<th>Concept 1 Preferred Name</th>
<th>Concept ID</th>
<th>Concept 2 Preferred Name</th>
<th>% Occurrence in Abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C0086969</td>
<td>Self-Management</td>
<td>C1171369</td>
<td>Adherence</td>
<td>36.6%</td>
</tr>
<tr>
<td>2</td>
<td>C1171369</td>
<td>Adherence</td>
<td>C2986593</td>
<td>Tailored Intervention</td>
<td>18.3%</td>
</tr>
<tr>
<td>3</td>
<td>C0030688</td>
<td>Patient education</td>
<td>C1171369</td>
<td>Adherence</td>
<td>16.9%</td>
</tr>
<tr>
<td>4</td>
<td>C1171369</td>
<td>Adherence</td>
<td>C3508152</td>
<td>Patient Engagement</td>
<td>16.9%</td>
</tr>
<tr>
<td>5</td>
<td>C0237125</td>
<td>Medication regimen</td>
<td>C1171369</td>
<td>Adherence</td>
<td>16.2%</td>
</tr>
<tr>
<td>6</td>
<td>C0085519</td>
<td>Reminder Systems</td>
<td>C1171369</td>
<td>Adherence</td>
<td>14.8%</td>
</tr>
<tr>
<td>7</td>
<td>C0162648</td>
<td>Telemedicine</td>
<td>C1171369</td>
<td>Adherence</td>
<td>13.4%</td>
</tr>
<tr>
<td>8</td>
<td>C0814098</td>
<td>health-related beliefs</td>
<td>C1171369</td>
<td>Adherence</td>
<td>13.4%</td>
</tr>
<tr>
<td>9</td>
<td>C0037438</td>
<td>Social support</td>
<td>C1171369</td>
<td>Adherence</td>
<td>12.7%</td>
</tr>
<tr>
<td>10</td>
<td>C0600564</td>
<td>Self Efficacy</td>
<td>C1171369</td>
<td>Adherence</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

Figure 1. Network graph of concepts identified from literature abstracts on drug therapy problems. The graph has a total of 50 nodes and 102 edges. Nodes represent UMLS concepts. Larger nodes have higher weighted degree centrality. Edges represent the frequency of co-occurrence of UMLS concept pairs across the document corpus. Thicker edges correspond to co-occurrence frequency. The top 10 nodes and edges, based on weighted degree centrality, are highlighted in blue.
Figure 2. The Enhanced Medication Therapy Management Evaluation Framework. Structure entities are shown in blue, process entities in orange and outcomes in green. Arrows indicate fluidity between structure, process and outcome constructs.
one-to-many mapping of text tokens including stop words to multiple concepts. Medication adherence was the most prevalent concept in the corpus, occurring in 93% of the abstracts. Other high-frequency concepts include self-management, tailored intervention and patient engagement (Table 3). These highly frequent concepts also formed the most frequent concept-concept pairs in the corpus (Table 4).

The network graph identified and analyzed in the study had 50 nodes and 102 edges (Figure 1). It is a sparse undirected graph with a density of 8.3%. The most dominant nodes and edges in the graph are highlighted in blue and further described in Tables 3 and 4 respectively. Medication adherence, connected to 47 nodes, is the most central node in the network. 19 nodes (38% of the full graph) are one-degree vertices connected to the medication adherence node. Its removal from the network results in a much smaller graph with 32 nodes and 55 edges. The other dominant nodes are closely related to the medication adherence as contributors or interventions to medication non-adherence. Interestingly, no other drug-related categories were identified in the network. The adverse drug event concept identified in the network analysis was discussed in the reviewed abstracts as an antecedent contributor to medication non-adherence rather than a drug therapy problem per se.

Conceptual Model Development

The enhanced model developed in this study meaningfully organizes the synthesized information in a manner that clearly shows the relationships and fluidity of the constructs and concepts about the application of medication therapy management for addressing drug therapy problems as illustrated in figure 2. The model consists of 65 concepts clustered into 14 constructs that are further organized into structural, process and outcome constructs. Structural constructs are shown in blue, process constructs in orange and outcome constructs in green. The constructs are derived from the parent theories and represent clusters of concepts that may contribute to drug therapy problems that were identified by network analysis and abstract review. The process constructs represent actors, activities, drug therapy problems and critical measurements associated with MTM service provision. The outcome constructs represent the key process, intermediate and clinical outcomes related to MTM service delivery. Because the concepts in this domain are highly heterogeneous, specific concept to concept relationships are not shown, and concepts have not been reduced to measurable variables.

Discussion

This paper describes the first iteration of a process used to develop a conceptual framework to guide a program of research in medication therapy management (MTM). We selected the MTM Service Model in pharmacy practice 3 and the Chronic Care Model 14 as parent theories and enhanced these through theory derivation 8. Content included in the enhanced model were extracted from abstracts of review articles using natural language processing and analyzed using network theory. The derived model was organized according to the structure, process and outcome constructs of the Donabedian Model 20. The framework requires additional elaboration as specific concept-concept relationships are currently not described.

Medication adherence was the most studied drug therapy problem. This may be attributed to the fact that medication adherence is often considered the single most significant predictor of treatment success for chronic diseases 21. Indeed, an estimated 70% of hospital admissions due to drug therapy problems are related to medication adherence 22. Our findings confirm the fact that adherence is a complex phenomenon influenced by a variety of heterogeneous factors, and that there appears to be no consensus on the most practical approaches for measuring and promoting adherence in chronic diseases 23-25. However, providing patient-centered care and enhancing patient self-management through patient engagement, education, and psychosocial support are accepted strategies for improving adherence.

Our review and model development highlights several knowledge gaps in the literature. First, clearer definitions of factors associated with drug therapy problems are needed. This would help researchers to formulate research questions and evaluate relationships in this domain more effectively. Second, there are knowledge gaps about factors associated with medication indication, effectiveness, and safety. More research on these drug therapy problems is needed to demonstrate their importance and the role of MTM in addressing them. Third, researchers have not adequately examined the process of task redistribution in the provision of MTM services. Task redistribution is the appropriate delegation of health service responsibilities to less specialized but more readily available health professionals, and in some cases non-professionals 26-29. Studies have shown that task-redistribution enhances patient engagement and self-efficacy through education and psychosocial support across different populations and intervention modalities 30-32. However, its role in MTM service provision remains unstudied. In high-resource settings where MTM services are primarily provided by pharmacists, the role of task-redistribution in MTM services is not adequately described.
Conversely, in resource-poor settings where task redistribution is popular, little is known of the standard practice of MTM. The application of mobile health (mHealth) technology to complement this approach by improving recording keeping and decision support has been shown to increase the quality and efficiency of this strategy.\textsuperscript{33,34} Lastly, while acknowledging the importance of complementing drug therapy problem intervention strategies with reminder tools and decision to support adherence,\textsuperscript{35,36} further informatics research is needed on the role these technology solutions in operationalizing of MTM service activities such as care coordination and eligibility determination.

Two key challenges should be expected when using natural language processing and network analysis to extract automatically and characterize concepts during conceptual framework development. First, the effectiveness of this approach is limited by currently available natural language processing technologies. For example, while natural language processing tools such as MetaMap can effectively automate the extraction of concepts from biomedical texts, technologies that could effectively extract semantic relationships between concepts are currently available. It is, therefore, necessary to rely on surrogate approaches such as pairwise concept co-occurrence analysis to identifying and quantify the strength of plausible relationships between concepts reported in biomedical texts. Second, natural language processing tools such as MetaMap often map idiosyncratic texts including stop words, acronyms, and abbreviations to concepts.\textsuperscript{12} Therefore, it may be necessary to review the extracted concepts to eliminate those that are not relevant. Because this review is manual, it may not be useful to scale the approach to larger numbers of biomedical texts.

There are several limitations in this study. First, due to pragmatic constraints associated with using MetaMap, we restricted our concept extraction to published abstracts of systematic reviews and meta-analyses indexed in Ovid Medline. This was done with the assumption that theoretical saturation would be achieved as long as a critical mass of literature texts were analyzed. Using full texts, including other types of published articles or incorporating articles from other bibliographic databases could have resulted in richer content, but would imply additional computation and manual review. Second, we relied on pairwise concept co-occurrence analysis as a surrogate measure of the relationships between concepts occurring in the analyzed abstracts. However, doing so does not give an accurate interpretation of the true strength of the relationship between co-occurring concepts in the analyzed abstracts. For example, a pair of concepts that occur in the same abstract may not be related. Conversely, just because two concepts do not co-occur together does not imply that they are not related directly or indirectly. We lost information about the directionality of plausible concept-concept relationships. Nonetheless, our goal at this stage was not to establish the true relationships between concepts, but to quantify what biomedical researchers have been interested in studying within the MTM domain. Our next steps are to refine the framework further by characterizing the interrelationships among the individual concepts and describing the measurable variables associated with these concepts. This will be followed by the use of domain experts to evaluate the framework extensively to determine its relevance across a variety of settings including in underserved settings.

Conclusions

This paper describes the use of natural language processing and network analysis for the development of a conceptual framework to guide a program of research in medication therapy management. The developed model consists of 65 concepts clustered into 14 constructs derived from the MTM service model and the chronic care model. It requires further refinement and evaluation to determine its applicability across a broad audience including underserved settings. Our analysis confirms that medication adherence is the most studied drug therapy problem. Patient-centered care approaches and enhancing patient self-management through engagement, education, regimen simplification and psychosocial support are important strategies for promoting adherence. Drug therapy problems associated with medication indication, safety and effectiveness require more studies. Further research is also needed to understand the roles of task redistribution and medical informatics interventions in operationalizing MTM activities. We recommend that researchers evaluating tools to address drug therapy problems should consider the fact that the efficacy of such tools could be influenced by a multitude of complex factors. In situations where controlled experiments cannot be carried out, these factors should be investigated as potential confounders and effect measure modifiers.

ACKNOWLEDGEMENTS

This study was not funded but was made possible by the Fulbright Science & Technology and the Columbia University Honorific Predoctoral Fellowships awarded to William Ogallo in pursuit of his Ph.D. degree in Biomedical Informatics.
References

Abstract
Health reminders are integral to self-managing chronic illness. However, to act on these health reminders, patients face many challenges, such as lack of motivation and ability to perform health tasks. As a result, patients experience negative consequences for their health. To investigate the design of health reminders that persuade patients to take action, we conducted six participatory design sessions with two cohorts: mothers of children with asthma, and older adults with type 2 diabetes. Participants used collages, storyboards, and photos to express design ideas for future health reminder systems. From their design artifacts, we identified four types of persuasive reminders for health self-management: introspective, socially supportive, adaptive, and symbolic. We contribute insights into desired features for persuasive reminder systems from the perspectives of patients and informal caregivers, including features that support users to understand why and how to complete health tasks ahead of time, and affordances for intra-familial and patient-provider collaboration.

Introduction
Reminders for health tasks, such as obtaining recommended lab tests and taking medications, are important for supporting patients who are living with chronic health conditions to manage their health. For these individuals and their families, reminders are a part of everyday life, helping patients to manage their health in the home rather than in the clinic or the hospital. Moreover, unlike acute care reminders for short-term health concerns, reminders for chronic illness care are often associated with making and maintaining significant changes in health behaviors, such as eating, exercising, and tracking health indicators over time. For these patients, simply alerting them to health tasks could be insufficient for supporting them to adopt and maintain medical and lifestyle changes. Rather, these individuals might benefit from reminders that are designed to be motivational, or even persuasive.

Strategies for supporting patients to change health behaviors and complete health tasks are well documented. In particular, increasing patients’ self-efficacy to solve problems through informational, emotional, and social support are key for catalyzing sustainable behavior change for health. However, the use of these strategies to design reminder systems for preventive and chronic health tasks is underexplored. Health reminders are typically designed to trigger health behaviors at the right time, but are not usually designed to employ persuasive strategies that can influence patients to adopt those behaviors. As a result, institutionalized reminders such as those coming from healthcare organizations likely do not live up to their full potential to support and encourage optimal self-care. Improving these reminder systems could enhance patient care, reduce disease burden, and potentially save health care costs for both patients and health insurers.

To inform the design of persuasive health care reminders, we explored desired persuasive strategies from the perspectives of individuals who are responsible for managing a chronic illness. Our goal was to understand how reminders could help patients to build self-management skills and engage in self-management tasks. We conducted six participatory design workshops with older adults who have type 2 diabetes and with mothers caring for children with asthma. Participants reflected on current reminder systems and designed future reminder systems that made it easier and more attractive to complete health self-management tasks. We found that the participants envisioned four types of persuasive health reminders: introspective, socially supportive, adaptive, and symbolic.

In this paper, we contribute empirical findings to inform the design of persuasive reminder systems based on the perspectives and expertise of patients managing chronic illnesses. We describe how persuasive health reminders are deeply personal, reflecting patients’ health values, relationships, and goals.

Related Work
In this section, we discuss the background of persuasive technology design, including persuasive design frameworks, strategies, and ethics. We focus on the application of persuasive systems in the health domain, but acknowledge that persuasive technologies have been applied in other domains such as sustainability and safety.
Persuasive technology design

Persuasive technology design emerged in the 1990s, and the health domain was identified early on as a key domain for such persuasive systems. Several models for designing persuasive technologies have informed the design of systems that influence health behavior change. One of the best known models is Fogg’s Behavior Model. Fogg’s model emphasizes three factors that must be present at the same time to influence behavior change: (1) motivation, (2) ability, and (3) triggers (e.g., alarms). When people have both the motivation and ability to perform a behavior, reminders that trigger the behavior at the right time are more likely to be effective. Typical health reminders aim to trigger tasks at the right moment. However, when lacking ability and/or motivation, people can benefit from triggers that facilitate and motivate them to perform a task. Fogg calls these triggers “facilitators” and “sparks,” and differentiates them from “signals” that remind rather than persuade. We used Fogg’s model to understand the significance of elements in participants’ designs that increased both their motivation and ability to perform health tasks and achieve health goals.

Building on these behavior change models, persuasive systems designers often employ specific, empirically grounded persuasive strategies, including: feedback about performance, rewards, authority, progress monitoring, and social influence. These persuasive strategies have been implemented in the design of systems that motivate health behavior change, including increasing physical activity, water intake, and engagement in diabetes self-care. These strategies have also been employed to design persuasive health reminder systems. For example, Walji et al. used iterative design to explore the acceptability of persuasive appointment reminders using the strategies of commitment, liking, authority, and scarcity. Another example is Langrial et al.’s email reminder system that used feedback and social support to increase logging of fizzy drink consumption. Finally, Oliveira et al. created a socially competitive game to engage elders in medication adherence, and found that it improved the timing and accuracy of taking medications. These examples highlight approaches to persuading people to perform recommended health tasks. Our research adds insights into the persuasive ways in which individuals link health tasks to their broader health goals for self-care.

Health behavior change researchers have also explored the design of personalized persuasive systems for health. For example, Kaptein et al. designed a reminder system to reduce snacking that sent reminders tailored to each user’s susceptibility to four social influence strategies: scarcity, authority, consensus, and commitment. They found that tailored persuasive reminders were more effective than random reminders. In another study, Kaptein et al. showed that persuasive reminders designed to increase physical activity and fruit intake were more effective when adapted to individuals’ persuasion profiles. In contrast to personalized persuasive systems that targeted a prescribed behavior (e.g., physical activity), Baumer et al. supported the agency of users to define healthy behaviors. Their system, VERA, enabled users to post photos of their health behaviors to encourage persuasive open-ended social awareness. Our findings show that there may be a middle ground between prescriptive and open-ended persuasive systems wherein users can contextualize prescribed health tasks within personally significant health goals.

Ethics of persuasive system design

Finally, the user’s agency in defining the strategies and goals of persuasive systems is an important ethical consideration. Davis described how design methods can influence the ethics of persuasive systems. She advocated for the use of Participatory Design (PD), a method that emphasizes equal participation and mutual learning between designers and users. PD is a human-centered design tradition that emphasizes the expertise of users, especially when designers have little first-hand knowledge of users’ experiences. For example, PD has been used to design with and for cancer patients, adults with amnesia, and children. Ethical dilemmas in persuasive systems can arise from the asymmetrical power relations between the persuader (i.e. the designer) and the user. PD is well-suited to balancing power relations, and increasing the user’s voice in the ways in which they are persuaded.

Our work contributes to this tradition of using PD to inform the design of ethical persuasive systems for health. In particular, we contribute insights into designing health reminders that align with patient’s personal health values and attitudes and that increase their confidence as self-managers. We identified four types of health reminders designed by patients to increase their motivation and ability to adopt health behaviors: retrospective, socially supportive, adaptive, and symbolic. These persuasive reminder types can inform new approaches to designing patient-centered systems that support health self-management.
**Participatory Design Method**

The participatory design approach enabled us to engage patients and caregivers in designing future reminder systems that could support them to manage chronic illnesses. Participants used creative, hands-on design activities to embody their perspectives in design artifacts, including collages, storyboards, and cultural probes. These artifacts enabled us to capture participants’ visions of future states of health and reminders.

**Participants**

We sampled 23 participants from two populations of people who managed complex chronic illnesses: older adults with type 2 diabetes, and mothers of children with asthma. This sampling strategy enabled us to include the perspectives of both patients and caregivers involved in chronic illness management and a wide spectrum of ages (31 to 89, µ=57). Within these two distinct populations, we oversampled individuals from underrepresented minorities and low socioeconomic backgrounds to ensure a diversity of patient and caregiver experiences on the design teams [see Table 1]. Participants were recruited from Group Health Cooperative, an integrated care and delivery system in Washington State.

We grouped participants into three design teams, each with 8 participants: (1) mothers of children with asthma (A1-A8); (2) older adults with diabetes (D1-D8); and (3) a mixed team of mothers and older adults (A9-A11; D9-D12). Each team participated in two 90-minute workshops, one week apart. Two researchers (the first and second authors) facilitated each workshop. Participants did three types of activities: collages, storyboards, and cultural probes.

**Table 1. Participant demographics.**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Age</th>
<th>Race/Ethnicity</th>
<th>Education Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mothers of children with Asthma (A)</td>
<td>11</td>
<td>31-45</td>
<td>AA A H W HS SC</td>
<td>4YR +4YR</td>
</tr>
<tr>
<td>Older adults with Diabetes (D)</td>
<td>12</td>
<td>54-89</td>
<td>73</td>
<td>6 0 0 6 4 2 2 4</td>
</tr>
</tbody>
</table>

AA=African American; A=Asian; H=Hispanic; W=White; HS=High School; SC=Some College; 4YR=4-year degree; +4YR= more than 4-year degree.

**Collages**

Collages are an ideal tool for helping people to “articulate experiences” using stimuli like images and words. We provided participants with a set of images relevant to chronic illness management, such as health care providers, patients, medical equipment (e.g. glucose monitors and asthma inhalers), hospital and home spaces (e.g., clinic office, bedroom, kitchen), wellness activities (e.g., walking, aerobics), and reminder tools (e.g., mobile phones, sticky notes, email). We also provided materials for free-form expression, including pens, colored paper, scissors, glue, and health-related magazines. The research team pilot tested this collage toolkit to ensure that it provided a balance of structure and ambiguity to encourage creativity.

Participants created two collages. In Design Workshop 1, participants worked independently to create a collage representing a magical system that would help them to achieve a health goal, such as healthy blood glucose levels, weight loss, or asthma control. In Design Workshop 2, we handed participants their collages and toolkit from the first workshop, and invited them to build off of these to imagine explicitly how reminders for health tasks fit into the picture of health self-management. Participants worked in pairs for the second workshop.

**Cultural probes**

Cultural probes are take-home materials that provoke participants to reflect on and capture aspects of their everyday environments—a common technique for stimulating design ideation. In between the first and second workshops, participants completed short activities at home about reminders that had persuaded them, and that they had authored to persuade someone else. We also provided each participant with a Polaroid camera to take photos of personally significant health reminders. The cameras were loaded with 5 mini films for instantly printing color photos.

**Storyboards**

In the second design workshop, after the paired collage activity, we asked participants to fill-in-the-blanks in a storyboard template. Storyboards and scenarios are common techniques for engaging users in the design process. We provided participants with a 3-panel storyboard depicting a patient “before” and “after” the intervention of a magical genie. In the first panel, participants wrote their thoughts and feelings when they experience a challenging health task. In the middle, they wrote what the genie would say or do to help them overcome this challenge. Then, in
the final panel they wrote their thoughts and feelings after the genie’s intervention. The focus on thoughts and feelings gave us insight into the psychological and informational barriers people faced when completing health tasks, and the type and tone of support they desired for successful task completion.

Analysis
Each workshop was video recorded and transcribed by an attending court recorder. The researchers analyzed the transcriptions, collages, storyboards, and cultural probe photos using a general inductive approach. The general inductive approach is a systematic way to develop a framework of underlying themes that connect clearly to research objectives. Our analysis was guided by the research goal of understanding the ways in which reminders helped people to achieve their health tasks and goals. The first two authors iterated on a codebook, first establishing themes such as, “rewards of health,” and then clustering these into broader categories such as “symbolic reminder.”

Findings
Participants designed persuasive health reminder systems aimed at provoking reflection on health priorities, increasing confidence in completing health tasks, tailoring the reminder mode and message, and symbolizing the rewards of good health. Below, we describe four types of persuasive health reminders envisioned by participants: introspective, socially supportive, adaptive, and symbolic.

Introspective reminders
Introspective reminders triggered reflection on health goals. These reminders helped patients plan and prioritize their daily health tasks. A6 created a collage that centered on taking time to reflect (Figure 1a). She described that, “I know in order for me to be successful, sometimes I have to sit back and reflect on those [health] goals. You know, what I can do to keep me going towards my health. So that's why I had the word reflect.” Her magical reminder was a kiosk at her front door where she could reflect on her goals and tasks: “I call it a self-check in. […] I would have to do a check in [to get] daily reminders or affirmations to get you going, or remind yourself what you need to work on today.”

Figure 1. Introspective reminders. (a) self-check in reminder; (b) health road map; (c) weight loss visualization.

A4 also described a reminder system that helped her to reflect throughout the day on healthful behaviors. “So my magical system, is just kind of about balance and time […] It's hard to maintain balance with other elements in my life. Like the exercise element. The healthy diet element. Relaxation and sleep element. […] It would be great to have some type of a system to help interface with all those different elements.” She phrased her reminders as questions that helped her to introspect on healthy behaviors: “Have I done enough stuff throughout the day? Have I eaten what I need to eat? Have I maintained enough sleep?” She explained why these types of reminders would be valuable, “When I get caught up in my work schedule, and go, go, go, do, do, do, sometimes those other elements in my life fall by the wayside and those are important for health too.”

Similar to A4’s daily, short-term introspective reminders, D7 and D8 designed a reminder system that supported self-monitoring over the short and long term. D7 and D8 worked together on their collage in the second workshop (Figure 1b). D8 described the introspective reminder system they created: “We liked the idea of having a health goals kind of road map, where there's short and long term goals, and that somehow you have access to that information. And incorporated in it are like little daily, hourly, encouragements, and humor, and things like that.”
D5 also created a reminder system that supported awareness of past, present, and future states of health (Figure 1c). She explained the nested colors depicting her losing weight over time, "So this [orange] is as I see myself now. This [green] is as I continue to work on these things. The smaller circles [yellow] are where I'm going to be—is where I'm going to." She pointed to the "1lb" water drops around her image, “And these are the pounds just dripping off me as I do exercise; and these are some of the things that I do to help, you know, make that happen.” This creative depiction of her health goal of “losing weight” shows the power of reminders that encourage reflection on desired future states of health. Participants described how these introspective reminders, featuring self-check ins, health elements, and road maps, would support them to connect their concrete health tasks to broader health goals.

**Socially supportive reminders**

In contrast to introspective reminders designed primarily for individual contemplation, **socially supportive reminders** involved motivational and mentoring relationships. Participants described persuasive health reminders that leveraged social support to not only remember tasks, but to increase confidence in performing them. These reminders were targeted at concrete health tasks like medication and exercise, rather than broader health goals.

One mother, A2, photographed a socially supportive reminder system that she was currently using to motivate her five children to take on more responsibility for their health. She described her system of collecting tally marks on the fridge to indicate a completed task, such as ‘Don’t forget to take your medicine’: “And for every time you do something you get a tally mark. And the goal is on the day that we score the highest score as a family we get to do something fun as a family. So everyone is working to try to get more tallies.” This system helped the mother to shift accountability for health self-management to her children, which benefited the entire family. This approach highlights the importance of designing health reminders for family units to change their health behaviors together. The mother added, “It’s a good way because then they’re reminding me. ‘We need to take our medicine because we want our points.’”

Many of the socially supportive reminders were exemplified in the storyboard activity. Participants were asked to depict how a magical genie would help them to overcome a health challenge. In her storyboard, D3 described the kind of social support that would help her to overcome obstacles to taking her blood sugar,

> “I was frustrated because I didn't want to take my blood sugars. I was upset because I wasted the needles, I poked my finger and the blood didn't come out and the machine said it wouldn't work. I didn't want to do it in the first place anyway, and so I was discouraged and upset. The magic person came in, showed me why it was helpful and how to be able to do it correctly. She infused in me confidence. She had an interest in knowing what the outcome was going to be and why it was important for me to be able to do that, and she infused in me the fact that it would be a great activity. After she did that, I felt much lighter. I also felt more energetic, positive, ready to get started.”

Being reminded of the rationale for the task, the outcome expectations, and her capacity to complete it, were central to this participant’s vision of a system that would increase her confidence to perform a self-administered blood glucose test. Some participant’s socially supportive genies were less comforting, yet still instructive. For example, D6 envisioned a genie who would help her to set up a reminder system that would prevent her from forgetting to take evening medications,

> “So genie said, ‘Well, you like little reminders. Put a reminder in an obvious place.’ And she said, ‘The solutions are put one in the bathroom, put one in the bedroom, put one wherever you're going to be before going to bed.’ So I put a sign on the mirror in the bathroom because I brush my teeth. And I usually have a cup of tea or something in the kitchen, so on the television I put a little small sign in red saying ‘Meds.’ So if I picked the cup up and looked at the screen, I would see the sign. And I said, ‘Thank you, genie,’ and genie said, ‘Think on your own.’”

This reminder was not a just-in-time alert to the task, rather it was well ahead of time and targeted at how to remember the task. Moreover, the “tough love” tone of social support was designed to encourage independence from the system. D2 also emphasized the importance of being reminded of how to refill medications so that he could ultimately take care of it himself,

> “I don’t have a car, so to get my medications at the clinic is kind of a chore. I have to go on the bus. So a genie came along and offered to fly me there. He also told me that they do have a system [at Group Health] for medication, but
you can't wait to the last minute anymore. Just got to plan ahead and order ahead of time. So at the end I was thankful for the genie, but also I realized I could take care of this problem myself.”

This reminder helped D2 to understand the time horizon for medication refills using the online system, which was essential to his a success with the task. These examples of socially supportive reminders demonstrate the importance of receiving certain reminders well ahead of time that explain why and how to complete health tasks.

Adaptive reminders
Participants designed adaptive health reminders that persuaded them by adapting to their personal health information. Some participants distinguished these personalized health reminders from standardized reminder letters provided by the health cooperative. For example, A2 and A3 worked together on a collage that emphasized this distinction. A2 explained, “We all get that form letter every year that says, ‘Hey, come in, your child is due.’ And it’s that generic form letter that’s been the same for the last twenty years or whatever and you don’t even read it anymore, do you? Like, ‘Yep, whoosh.’” The health tasks included in these standardized reminder letters are based on health conditions, gender, and age. Participants described how personalized letters would compel them to retain the information and act on it, as A4 said, “It makes me focus on the reminder. I will remember it a lot more and be willing to follow through. [...] I’m not just a number.” A1 suggested that, “maybe it would motivate us to get our health assessments done if we knew that we were going to get more personalized tips.”

Participants highlighted the importance of using tailored reminders to catch their attention and help them to prioritize health tasks. For example, A3 designed a system that anticipated health needs and made suggestions based on her and her children’s personal health information. She explained, “This one would be a super app, so it would have your medical goals and specifications. Like my daughter’s asthma would be on there. And then I thought it would be cool if it helped you maintain an active lifestyle. If you’re going to exercise, it asks you: ‘How much time do you have to exercise today?’” She continued, “It might give you guidelines for how much sleep your kids need for their age group.” This example suggests that persuasive systems will use personal health information to tailor reminders that help people to choose highly relevant healthy behaviors. Moreover, designing for the family unit is an important feature of reminder systems envisioned by caregivers.

We also found that the personalization of health reminders applied to both the medium (i.e., the technology) and the message (i.e., the content) of the reminder. For example, in contrast to A3’s design of a mobile phone application, D3 and D4 designed a reminder system for diabetes self-management that was embedded in a refrigerator—a device that was very central in their chronic illness. The system was designed to adapt health reminders and notifications based on data that it collects from the user. D3 explained, “The refrigerator would see what his [glucose] numbers are, be concerned about the fact that his number is low or high, and would be able to tell him that his number is low, [and] he should probably not eat certain type of things.” The system also used the personal data to help the patient prepare for appointments. D3 continued, “It reminds you, ‘This is your appointment; this is your appointment day; this is the time; and these are the things that you need to discuss with the doctor,’ based on the things that you’ve already inputted into the thing.” This example highlights the persuasiveness of embodying reminders in a medium that is relevant to the task, in this case, the refrigerator is directly relevant to healthy eating.

D1 and D2 created another example of a centrally located technology designed to gather personal health information. Similar to D3 and D4, their system highlighted the importance of enabling patients to coordinate care with their providers. They designed an information “hub” that automatically generated appointments and reminders in response to data inputted by patients and their providers. This system would enable patients to shift accountability for certain health tasks to their providers. For example, rather than patients initiating health appointments and reminders, this system helped providers to monitor health data so that they could create appointments and reminders as needed. D1 explained that after the provider inputs lab test results,

“It goes back to the patient; and the patient, when we do all of our testing at home, you can put that in your computer, too [...] So all of that [personal data] will be in there; and it can help the doctor tell you when the next appointment is, when the next test is. And all of that goes into this one computer program which collects it all and generates it [reminder].”
Adaptive reminders revealed patients’ desire for shared accountability among provider, system, and patient. To simplify and engage with complex self-management tasks, patients designed reminder systems that reduced the burden of personalization while facilitating active patient-provider collaboration.

**Symbolic reminders**

The final type of reminder we identified was **symbolic reminders**. Symbolic reminders motivated participants to do health tasks by reminding them of personally significant reasons for healthy behaviors. These reminders were exemplified in the cultural probes data, wherein participants were instructed to take photos of helpful reminders using a Polaroid camera. Their photos were often deeply personal, reflecting the relationships and values that helped them to put health in perspective. For example, A4 featured her son as her reminder (Fig. 2a), “I took a picture of my son and I outside of the YMCA here, so it’s just a reminder to me to keep fit and healthy and active.” Similarly, D1 photographed her dog (Fig. 2b) who reminded her to walk every day, “My puppy gets very annoyed if we don’t go walking every day, at least once, so that’s why I took a picture of her [...] She follows me around and I say, ‘Do you want to go for a walk?’ And she runs to the door. So, anyway, she’s a reminder.”

**Figure 2. Symbolic reminders.** (a) Child; (b) Dog; (c) Flowers; (d) Medications.

Some participants photographed symbols of the rewards of good health. For example, D7 described a reminder that symbolized enjoyment of the outdoors. He used his Polaroid camera to take “a picture of being able to look outside and enjoy the day. [...] Just a reminder to myself, I’ll get a chance to see another day.” D9 also took a picture of the outdoors; hers focused on flowers (Fig. 2c). She explained how she was temporarily barred from sewing because of pain in her hand, “I’m off limits now with this hand. It’s driving me crazy, so I got to remember ‘I’ll get back to it, just relax.’ [...] If I can’t sew, I’ll go out and smell the flowers and take it easy.” In contrast to these positive symbols of good health, D6 reminded himself of the costs of poor health (Fig. 2d): “I took a picture of all of my medications because whenever I look at them I’m reminded of how much poor health can cost you. They all add up, and so they remind me to try to watch my weight and watch what I eat so hopefully keep my problems under control—hypertension and diabetes and also my chemotherapy drugs.” These symbolic reminders were part of patients’ everyday lives, motivating them to seek the rewards of healthy behaviors.

**Discussion**

Our participatory design approach to understanding persuasive reminders for health self-management revealed several implications for future systems. First, participants challenged the traditional conception of delivering reminders at the right time. Rather, participants emphasized the persuasive value of reminders that sparked behaviors well ahead of time. For example, introspective reminders were oriented toward helping patients to plan ahead and prioritize health tasks. Participants emphasized that being triggered well ahead of tasks was important for helping them to build awareness of their health trajectory and contextualize their health tasks within it. Introspection on the relationship between health tasks and goals enabled patients to reflect on cause-and-effect relationships between health behaviors and outcomes over long time spans. This reflection aided contemplation on the rationales—the why—for performing health tasks. Similarly, symbolic reminders that were embedded in everyday activities and relationships were not conceptualized as just-in-time triggers. Rather, the image of a child, a favorite
hobby, or an outdoor scene were omnipresent reminders that heightened awareness of personally significant rationales for being healthy. Thus, reminders that spark health behaviors well ahead of time enable persuasive messaging that increases self-awareness and motivation to perform health tasks.

We found that, in addition to systems that increased motivation, participants designed systems that increased their ability to perform them. While introspective and symbolic reminders helped patients to understand why they were doing health tasks, socially supportive and adaptive reminders emphasized how to perform tasks. Such systems provided them with information about resources, services, and instructions relevant to tasks that helped them to develop expertise. For example, some participants designed “divine interventions” by genies who reminded them of the skills and resources needed for tasks. The focus on skills mastery that characterized socially supportive reminders was designed to facilitate shifting accountability for health tasks from parents to children, and from patients alone to patient-provider collaboration. Similarly, the emphasis of adaptive reminders on anticipating needs and making recommendations created new opportunities for sharing accountability between patients and providers. Participants valued socially supportive and adaptive reminders for enabling them to reduce the burden and frustration of chronic illness management. They described how these reminder systems would help them have more confidence in their ability to be effective self-managers and to work together to remember and achieve health goals.

We also found that participants considered personalization to be crucial to persuasive reminders for health self-management. Participants linked personalization to feeling more engaged in health care and more willing to follow through on health tasks. They designed systems that were personalized in terms of both the message being delivered, and the mode of its delivery. For example, a mother designed a mobile-phone based application that delivered tailored messages about asthma care based on her child’s personal health information. The application would detect weather conditions and other information relevant to asthma to create reminders as the context changed. An older adult designed a refrigerator-based system that gathered information and delivered reminders relevant to nutritional management for patients with diabetes. These examples demonstrate that the technological embodiment of reminders for asthma require different affordances than those for diabetes care due to the unique contexts of certain health behaviors. Studies have demonstrated that persuasive health reminders personalized to individuals’ susceptibility to different persuasive strategies exceed the impact of generic reminders. Our findings confirm and expand this finding by suggesting that patients are most likely to be persuaded by health reminders that are personalized to their health goals and values, in addition to their preferences for different persuasive strategies.

Finally, the four types of persuasive reminders that we identified—introspective, socially supportive, adaptive, and symbolic—had different affordances for individual, intra-familial, and patient-provider interactions. These different patterns of interaction influenced both the scope of persuasive strategies, and the roles of patients, caregivers, and providers in health self-management. For example, introspective reminders were primarily designed for individual contemplation, whereas socially supportive reminders aided families, and adaptive reminders featured collaboration between providers and patients. Designing for these different units of analysis offers new ways to innovate health reminders that promote shared accountability and social learning for health. Our findings point to the importance of designing persuasive health reminder systems that individuals, families, and patient-provider pairs can interact with.

**Conclusion**

We used a participatory design approach to understanding patients’ and caregivers’ desires for health reminder systems that persuade them to manage the myriad and complex tasks of preventive and chronic illness management. Participants embodied their perspectives of desired futures in design artefacts that revealed four types of persuasive reminders for health self-management: introspective, socially supportive, adaptive, and symbolic. We contribute insights into implications for the design of persuasive health reminders including designing for the how and why of performing health tasks, triggering tasks well ahead of time, and personalizing both the message and the medium of reminders. Our contributions to persuasive system design are grounded in an ethical approach to engaging patients as authors of their own persuasive strategies that reflect their attitudes toward health self-management. This approach is significant for informing the design of patient-centered systems that empower people to adopt health behaviors in ways that are sensitive to their cultural, personal, and interpersonal values.
References


Assessments of the Veteran Medication Allergy Knowledge Gap and Potential Safety Improvements with the Veteran Health Information Exchange (VHIE)

Eric Pan, MD1,2, Nathan Botts, PhD1,2, Harmon Jordan, ScD1,2, Lois Olinger, MA1,2, Margaret Donahue, MD2, Nelson Hsing, ScD2

1Westat, Rockville, MD; 2US Department of Veterans Affairs, Washington, DC

Abstract

The U.S. Department of Veterans Affairs (VA) Veteran Health Information Exchange (VHIE, formerly Virtual Lifetime Electronic Record, or VLER) had been deployed at all VA sites and used to exchange clinical information with private sector healthcare partners nationally. This paper examined VHIE’s effect on allergy documentation. Review of all inbound VHIE transactions in FY14 showed that VHIE use was associated with a nearly eight-fold increase in allergy documentation rate. Preliminary manual document review further showed that VA and partners had shared knowledge of only 38% of patient allergies, while VA had exclusive knowledge of another 58% of patient allergies, and partners had exclusive knowledge of the last 5% of patient allergies. To our knowledge, this is the first study that examined the effect of HIE on allergy documentation.

Background

Allergies rank sixth as a cause of chronic illness in the United States, affecting more than 50 million Americans each year with an annual cost of more than $18 billion. Drug allergies worldwide are estimated to occur in 1% to 2% of all hospital admissions and 3% to 5% of hospitalized patients. Because allergic reactions may cause death or severe morbidity, documentation and exchange of allergy information is necessary for appropriate patient care. Missing information such as the knowledge that a patient has a life-threatening allergy to a medication—can result in catastrophic adverse events. In addition, maintaining medication allergy lists is a functionality included in the Centers for Medicare and Medicaid meaningful use program.

Interoperability across electronic health record (EHR) systems is important so that providers can share allergy information. Using a health information exchange (HIE) enables the storage of information about active allergies and medications so that providers can view medication lists and improve patient safety by avoiding drug/allergy interactions. Having drug information can improve efficiency by reducing the amount of time spent reviewing medication lists. Electronic medication administration records, ePrescribing, and provider order entry depend upon accurate medication and allergy lists. A U.S. Government Accountability Office study found that patient allergy information was among the most critical types of information needed by providers through health information exchange. That study reported that existing standards for exchange of allergy information are not adequate; one EHR system may label an allergic reaction as a side effect, while another system may classify the same reaction as an allergy.

Goss et al. indicated that the correct terminology for encoding allergies is complicated and that there are gaps in existing standards. They described the differences between the terms allergy, intolerance, hypersensitivity, and adverse sensitivity and pointed out that allergies are often poorly documented in EHRs. They conducted a comparative analysis of different standard terminologies for encoding information about allergies and recommended criteria for assessing standard terminologies. The Veterans Administration (VA) and the U.S. Department of Defense conducted early work in this area and constructed a framework to exchange standardized codified drug allergy information.

In a survey of emergency department physicians’ views of HIE, 21% of respondents chose allergy information as one of the top five data elements they would like to see included in an HIE. A survey of primary care physicians found that allergy information was described, among only three items, as highly important. Another survey found that most respondents indicated that verifying a medication and allergy list obtained through an HIE was faster than creating one.

Objective

The VA is pursuing interoperability between its EHR and private sector EHR systems through its Veteran Health Information Exchange program (VHIE, formerly known as the Virtual Lifetime Electronic Record, or VLER program.) Potential outcomes of VHIE include support for decision making, improved quality, safety and
coordination of care, reduced redundant testing and cost savings, and more and better access to information. The VHIE program has the potential to increase availability of allergy information to VA clinicians and partner clinicians in the private sector. The objective of this article is to assess whether VHIE helps improve allergy documentation in the EHRs of veterans seeking care by examining several different aspects of allergy documentation in different subgroups or contexts:

1. How often do VHIE-supported VA visits (i.e., a VA clinical encounter where the VA provider retrieved a VHIE document from an external partner organization) coincide with new allergy entries or updates to existing allergy entries? How does this rate compare with that of a non-VHIE Veteran?
2. How often are VHIE partner organizations aware of allergies in Veterans whom VA believed to have No Known Drug Allergies (NKDA)?
3. In Veterans who have documented allergies at VA, how do VA and partner allergy documentations compare? How complete are VA records? How complete are partner records?

Methods

We defined the intervention group as all Veterans who had an inbound VHIE transaction in fiscal year 2014 (FY14, October 2013 through September 2014) as reported by the Veteran Authorization and Preference (VAP) system, with all test users and test transactions removed. The VAP server was also used to identify all VHIE-participating Veterans so they can be excluded from the control group.

The Veteran demographic information reported by the VAP server was then linked to Veteran visit history and VA allergy documentation stored in VA's clinical data warehouse (CDW). SAS software was used to retrieve data via SQL queries and calculate the following variables:

1. What percentage of FY14 VHIE inbound transactions occur on the same day as a VA visit and either a) a new allergy record, or b) an update to an existing allergy record?
2. What percentage of 10,000 randomly selected FY14 VA visits in non-VHIE participating Veterans occur on the same day as either a) a new allergy record, or b) an update to an existing allergy record?

The authors then conducted preliminary manual record reviews using two separate samples of 10 charts:

1. Review of all inbound VHIE documents of 10 randomly selected Veterans in the intervention group who has no VA allergy records. The Veterans were then classified into three categories: a) Confirmed NKDA from partner organizations; b) No allergy documentation from partner organizations; or c) Documented allergies from partner organizations.
2. Review of all VA allergy records and all inbound VHIE documents of 10 randomly selected Veterans in the intervention group who had one or more VA allergy records. One of the physician authors (Dr. Pan) then manually reconciled and de-duplicated all documented allergies for each Veteran to determine a) total number of allergies documented for each Veteran; b) the number of allergies known only to VA; c) the number of allergies known only to partner organizations; and d) the number of allergies known to both VA and partners organizations.

The data analysis for this paper was generated using SAS software, Version 9.3 of the SAS System for Microsoft Windows and Microsoft Excel 2010.

Results

During FY14, over 6,000 Veterans benefitted from VA providers retrieving their non-VA medical records via VHIE. After de-duplicating multiple retrievals on the same day, there were 6,331 unique VHIE inbound transactions (i.e. unique Veteran-date pairs) in FY14. When cross-referenced against VA’s CDW allergy documentation database, there were 476 allergy creation or modification entries for these Veterans that occurred on the same day as their respective inbound VHIE transactions, or approximately 7.5% of the VHIE-supported VA visits (See Table 1: Frequency of Allergy Updates in VHIE and non-VHIE Visits below). In comparison, in a sample of 10,000 randomly selected visits from non-VHIE Veterans, only 84 visits were associated with a corresponding change in allergy documentation. Based on these comparisons, the use of VHIE is associated with a nearly 8 fold increase in allergy documentation rate (7.5% vs. 0.84%).
Table 1: Frequency of Allergy Updates in VHIE and non-VHIE Visits

<table>
<thead>
<tr>
<th>Count</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Veterans with inbound VHIE transactions in FY14</td>
<td>6,183</td>
</tr>
<tr>
<td>Unique VHIE inbound transactions in FY14</td>
<td>6,331</td>
</tr>
<tr>
<td>Inbound VHIE transactions on the same day as an allergy update</td>
<td>476/6,331 (7.5%)</td>
</tr>
<tr>
<td>Non-VHIE visits on the same day as an allergy update</td>
<td>84/10,000 (0.84%)</td>
</tr>
</tbody>
</table>

In addition to the analysis of allergy documentation rates discussed above, the authors manually reviewed the VHIE transactions and allergy records of 10 Veterans with no known drug allergy and 10 Veterans with known allergies. For the 10 Veterans in which VA is unaware of any drug allergies, 3 of them were explicitly documented as NKDA by VHIE partners. In addition, 6 of the Veterans simply had blank or empty allergy sections in their VHIE documents, passively confirming that the VHIE partners are also unaware of any drug allergies. However, VHIE partners reported a previously unknown drug allergy on one of the Veterans, an important patient safety contribution to VA sites caring for this Veteran (Figure 1, below).

Figure 1. Partner knowledge when VA has no documented allergies

The authors also reviewed 63 VA allergy records and 25 VHIE partner allergy records on 10 Veterans with known drug allergies (Figure 2. Number of allergies documented by VA and Partners, below.) After manual reconciliation and de-duplication, there were 53 unique Veteran allergies documented. VA allergy documentations included 49 out of these 53 unique Veteran allergies. As such, VA is aware of 95% of the Veteran’s allergies (Figure 3. Percentage of allergies known to VA vs. Partners, below.) Based on the VHIE documentations reviewed, VHIE partners are only reporting 25 unique Veteran allergies, or merely 43% of all known allergies. In addition, 38% of the 53 unique Veteran allergies are known to both VA and VHIE partners. So, only 5% of all allergies are known exclusively to VHIE partners.
Discussion

This paper presented a preliminary study of HIE’s impact on allergy documentation through analysis of the correlation between VA providers’ retrieval of external clinical documentation through VHIE and their updates to the Veteran’s allergy documentation. We further examined the contents of the shared VHIE allergy documentation to support the potential for VHIE information sharing to improve both VA and external clinician’s knowledge of Veteran allergies.

At baseline, our random sample of 10,000 VA visits showed an allergy documentation update rate of 86/10,000, or 0.86%. Considering that Veterans in our sample ranged from 1 visit per year to more than 100 visits during FY14, the update rate could be as high as 1 allergy entry update per Veteran per year, which is not insignificant or unreasonable. However, it paled in comparison to the allergy documentation update rate of 7.5% among VHIE-supported visits to VA providers. This was a 772% increase, or nearly an 8-fold difference. There’s several possible explanations for this dramatic difference. First, the VA allergy documentation database captured both new allergy entries and updates to existing allergy entries. So these transactions included both when VA providers learned of a new Veteran allergy from non-VA records, and when an existing allergy documented by VA is confirmed by external records. Second, we may be observing a transient increase in allergy documentations due to the initial
reconciliation between two or more data sources of Veteran allergies. This would imply that the allergy documentation rate would decrease, or decay, as more documents are exchanged on the same patient. Finally, some providers may simply be copying external allergy documentations into VA records en masse without deduplication for medicolegal purposes, leading to a superficial increase in allergy documentation rates. Therefore, future studies should follow allergy documentation rates in the same population over time to monitor whether this increase is sustainable. In addition, in-depth content analysis of existing VA allergy documentation, VA allergy documentation updates, and allergy documentation shared by VHIE partners would help to confirm whether the information shared and updated had clinical significance.

In the small group of ten Veterans with known drug allergies to VA examined in this study, less than 40% of allergies were known to both VA and external providers. This is a dangerous situation for patient safety. While we did not conduct further chart reviews to determine how much patient harm occurred due to this lack of information, some avoidable events undoubtedly occurred. This preliminary and rudimentary analysis highlighted the current knowledge gaps and illustrated the potential patient safety impacts of HIE efforts such as VHIE.

Although these results are preliminary and limited in scope and sample size, the three analyses demonstrated reasonable internal consistency. When comparing indicators of VHIE partner contribution of allergy documentation and knowledge, the VA provider documentation update rate of 7.5%, the partner report of allergy in 10% of Veterans with no known drug allergies to VA, and the 5% allergies known exclusively to partners, all help to constrain VHIE partner contribution to VA allergy knowledge between 5 to 10%. In comparison, VA’s ability to contribute 58% of unique knowledge about Veteran allergies cannot be understated, and should continue to motivate VA to promote sharing of VA records with non-VA providers.

**Conclusion**

To our knowledge, this is the first study to examine the relationship between HIE use and allergy documentation. In this sample of 6,183 Veterans with 6,331 inbound documents in FY14, the use of VHIE was associated with an eight-fold increase in the allergy documentation rate (7.5% vs. 0.84%). In addition, selected manual review of both Veterans with no known allergies and known allergies had shown that VHIE partners can contribute previously unknown (to VA) allergy information at least 5–10% of the times. It was informative to learn that VHIE partners are only aware of less than half of Veteran’s allergies. As such, continued efforts to promote partner use of VHIE to retrieve VA documentations would greatly enhance partner’s knowledge of Veteran allergies and improve patient safety.

**References**

Leveraging Terminology Services for Extract-Transform-Load Processes: A User-Centered Approach

Kevin J. Peterson, MS\(^1\), Guoqian Jiang, MD, PhD\(^2\), Scott M. Brue, BS\(^1\), Hongfang Liu, PhD\(^2\)

\(^1\) Division of Information Management and Analytics, Mayo Clinic, Rochester, MN
\(^2\) Department of Health Sciences Research, Mayo Clinic, Rochester, MN

ABSTRACT

Terminology services serve an important role supporting clinical and research applications, and underpin a diverse set of processes and use cases. Through standardization efforts, terminology service-to-system interactions can leverage well-defined interfaces and predictable integration patterns. Often, however, users interact more directly with terminologies, and no such blueprints are available for describing terminology service-to-user interactions. In this work, we explore the main architecture principles necessary to build a user-centered terminology system, using an Extract-Transform-Load process as our primary usage scenario. To analyze our architecture, we present a prototype implementation based on the Common Terminology Services 2 (CTS2) standard using the Patient-Centered Network of Learning Health Systems (LHSNet) project as a concrete use case. We perform a preliminary evaluation of our prototype architecture using three architectural quality attributes: interoperability, adaptability and usability. We find that a design-time focus on user needs, cognitive models, and existing patterns is essential to maximize system utility.

INTRODUCTION

Controlled terminologies are an important part of the medical record. The semantics they provide to heath care data is critical not only for interoperability between systems and institutions, but for analytics and secondary use[1]. Demand for these terminology resources at an enterprise level has driven the creation of functional requirements[2], formalizing the role of a terminology service in the enterprise. Because of their broad importance, terminology services are generally thought of as infrastructural components, existing to support other applications[3, 4] such as clinical decision support[5] and data processing pipelines[1].

There are, however, many instances where users may interact more directly with terminologies. One such example is the Extract-Transform-Load (ETL) use case. ETL scenarios, usually involved in data warehousing, not only must transform data structure, but data semantics as well[6]. Metadata describing the data semantics is not always readily available[7], which frequently places the responsibility of capturing, managing, and mapping these semantics on the ETL developers. This presents two problems. First, semantic transformation logic may become intermixed with the structural transformation process, and any captured terminology becomes a black box rather than an explicit, shareable knowledge artifact[8]. Second, clinical terminology is complex, and ETL developers often need subject matter experts’ and clinicians’ engagement to properly capture semantics[9].

In this work we present a design approach to the application of terminology services, centered around the ETL use case, that is deliberately focused on the needs and goals of the user, or user-centered[10]. We propose this not because existing terminology services are insufficient, but because domain mastery of terminology modeling and services is not universal, and should not be a prerequisite for utilizing these services – a notion we borrow from a recent study of bioimaging software[11]. Lowering the barrier to entry to these services while keeping users’ processes and cognitive models in mind increases usability[10] and ultimately, utility. Our focus is less around a terminology service in isolation, and more centered on a terminology system – or, the service along with other components in the context of a use case. With this in mind, we present a set of architectural principles necessary to satisfy both our functional and usability goals of a terminology system. Finally, through a concrete case-study, we analyze a prototype system against these principles.
MATERIALS AND METHODS

Use Case: The Patient-Centered Network of Learning Health Systems

The National Patient-Centered Outcomes Research Institute (PCORI), authorized by Congress in 2010, is an independent nonprofit, nongovernmental organization centered around Clinical Effectiveness Research (CER)[12] with the goal of advancing clinical care by utilizing an evidence-based comparison of methods[13]. The National Patient-Centered Clinical Research Network (PCORnet), a PCORI initiative, focuses on the development of networks to share health care related data[14]. The Patient-Centered Network of Learning Health Systems (LHSNet), a collaboration of ten institutions, is our specific segment of the PCORnet network[15]. The goals of LHSNet build on the objectives of PCORI and PCORnet, and focus on the ability to distribute queries throughout the network. Although the sites are not exchanging patient data – only summary-level query results – a precondition for the network is that each site maintain a common data model-based data mart.

PCORnet Common Data Model. It is necessary for any distributed data set to agree on common data structures and semantics[16]. For PCORI, this agreement is in the form of the PCORnet Common Data Model (CDM)[17]. The PCORnet CDM provides database agnostic information on the data structure (such as data types and referential integrity), as well as data semantics (in terms of allowable values for codified data).

Creating the Data Mart. Construction of a PCORnet CDM compliant data mart introduces our ETL use case. Integrating multiple diverse data sources into a common representation presents significant challenges. First, the data must be structurally aligned at a schema level. Given multiple data sources, this problem compounds[18] – naming of objects and attributes may differ, and differences in granularity will need to be reconciled. Structural differences are only one aspect, however. Semantic alignment needs to be considered as well, and is of particular interest to this study because it often requires a dedicated terminology service[19, 20, 7]. Semantic alignment, or the reconciliation of analogous codifications, is an important aspect to maximizing data utility, especially across institutions[21].

Related Use Cases. LHSNet draws many similarities to other multi-institutional research network initiatives – mainly, the Informatics for Integrating Biology and the Bedside (i2b2)[22] and Observational Health Data Sciences and Informatics (OHDSI)[23] projects. We explore below in brief the data model and terminology integration of the two.

The i2b2 data model is centered around observations on a patient, or “facts”[24]. Each of these facts is associated with a single concept – either a locally defined term or a concept from an external standard. The concepts themselves are arranged into a collection, or i2b2 ontology, which can then be used to build dynamic queries into the data.

OHDSI uses the Observational Medical Outcomes Partnership (OMOP) Common Data Model as a data specification. WhiteRabbit,1 an OHDSI developed tool, supports the ETL process by allowing examination of the source and target database structure and content. OMOP provides extensive guidance around both terminology and its structure via the OMOP Common Data Model Standard Vocabulary Specification Version 4.5[25]. Both of these use cases require the use of a defined CDM, and thus, require some sort of ETL process to import data. As with LHSNet, this data import requires the management of the data semantics. Huser and Cimino, while analyzing both OMOP and i2b2, suggest that “[a] terminology driven ETL process” would be a useful, especially when mapping local codes to external standards[20]. This observation aligns with our examination of the LHSNet use case, and suggests some degree of a shared need.

Terminology Models and Services

Value Sets. For our purposes, we define a value set as a set of codes used to define a semantic domain of interest. They must have some notion of a unique identifier and also may carry various metadata, such as names and descriptions. The PCORnet CDM defines 70 value sets to constrain attributes in the data model. These value sets are identified by a unique name and contain concepts specified by a code, name, and optional description. For example, the PCORnet CDM Race value set is shown in Table 1.

1Source code available at https://github.com/OHDSI/WhiteRabbit
<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>American Indian or Alaska</td>
<td>A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.</td>
</tr>
<tr>
<td></td>
<td>Native</td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>Asian</td>
<td>A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.</td>
</tr>
<tr>
<td>03</td>
<td>Black or African American</td>
<td>A person having origins in any of the black racial groups of Africa.</td>
</tr>
<tr>
<td>04</td>
<td>Native Hawaiian or Other</td>
<td>A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.</td>
</tr>
<tr>
<td></td>
<td>Pacific Islander</td>
<td></td>
</tr>
<tr>
<td>05</td>
<td>White</td>
<td>A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.</td>
</tr>
<tr>
<td>06</td>
<td>Multiple race</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>Refuse to answer</td>
<td></td>
</tr>
<tr>
<td>NI</td>
<td>No information</td>
<td></td>
</tr>
<tr>
<td>UN</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>OT</td>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: PCORnet Common Data Model v3.0 Race value set

Mappings. Mapping is a process of asserting relationships between concepts from different value sets[26]. These relationships can signify equivalence (or degrees thereof), hierarchical association (such as broader/narrower), or that no match exists for a concept. This is often done for interoperability purposes, where diverse data sets need to be aligned into a common semantic model[27]. It is also prevalent in ETL and data warehousing operations, where the target data model may specify value sets to which the incoming data must conform. To align these semantics, an institution will need to map their internal value sets to the target data sets. This is certainly the situation with our LHSNet use case, as multiple internal value sets need to be aligned to the PCORnet CDM semantics. Table 2 shows a sample view of a mapping of institution-specific codes to the PCORnet Race value set.

<table>
<thead>
<tr>
<th>Internal Code</th>
<th>Internal Name</th>
<th>PCORnet Code</th>
<th>PCORnet Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>American Indian/Alaskan Native</td>
<td>01</td>
<td>American Indian or Alaska Native</td>
</tr>
<tr>
<td>B</td>
<td>Black or African American</td>
<td>03</td>
<td>Black or African American</td>
</tr>
<tr>
<td>BA</td>
<td>African</td>
<td>03</td>
<td>Black or African American</td>
</tr>
<tr>
<td>BB</td>
<td>American Born African</td>
<td>03</td>
<td>Black or African American</td>
</tr>
<tr>
<td>BC</td>
<td>Caribbean Black</td>
<td>03</td>
<td>Black or African American</td>
</tr>
<tr>
<td>BM</td>
<td>African American</td>
<td>03</td>
<td>Black or African American</td>
</tr>
<tr>
<td>C</td>
<td>White</td>
<td>05</td>
<td>White</td>
</tr>
<tr>
<td>N</td>
<td>Native Hawaiian/Pacific Islander</td>
<td>04</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>NG</td>
<td>Guamanian or Chamorro</td>
<td>04</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>NH</td>
<td>Native Hawaiian</td>
<td>04</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>NO</td>
<td>Other Pacific Islander</td>
<td>04</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>NS</td>
<td>Samoan</td>
<td>04</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>P</td>
<td>Asian</td>
<td>02</td>
<td>Asian</td>
</tr>
<tr>
<td>O</td>
<td>Other</td>
<td>OT</td>
<td>Other</td>
</tr>
</tbody>
</table>

Table 2: A tabular representation of a semantic mapping between an internal Race value set and the PCORnet CDM v3.0 Race value set.
General System Functionality

Our goal is not to enumerate the functional requirements of our specific use case. We aim to reason about a broader architecture – one capable of supporting functionality outside of our current needs. There are, however, some general functional statements we can assume for our system that should broadly apply. Although not exhaustive, the following functional areas of concentration serve to scope our discussion.

**Terminology Content and Services.** Access to terminology resources depends on two factors: (1) the robustness of the structures used to model them, and (2) the completeness of the services in place to access these structures. At a minimum, the system shall provide a model for value sets and mappings, as well as services providing Create, Read, Update, and Delete operations. Content, both institution-specific and externally defined standardized terminologies, is generally necessary but use case dependent.

**Graphical User Interface.** User interfaces are often highly contextual, and effective ones focus on the specific concerns of the intended user base[28]. Because of this, we will not attempt to enumerate all user interface requirements here. At a minimum, manipulation of both value sets and mappings should be supported. Also, discussion and collaboration tooling is known to be important to the overall process of terminology refinement and validation[29], and should be supported natively to promote broader user engagement.

Architectural Quality Attributes

Evaluating the quality of a proposed architecture is a non-trivial task[30]. Quality goals often become represented as non-functional requirements, such as *the system shall be easy for users to use*. This not only presents implementers with a requirement that is untestable[31], it often pushes high-level goals down to individual components or modules, where the overall context can be lost. To avoid this, Architectural Quality Attributes[32] may be used to state system-level goals. They also serve to both influence and measure the system architecture. Below are three important quality attributes for our proposed system. This is by no means an exhaustive list, but was considered the minimal set to meet our stated system characteristics. Implementers are expected to add to this list as appropriate.

**Quality Attribute 1: Interoperability.** With this quality attribute, we aim to measure how well the components of our system share data with each other and external systems[33]. In previous work, the Strategic Health IT Advanced Research Projects (SHARP) demonstrated the need for standardization of data semantics using terminology services when normalizing Electronic Health Record (EHR) data[1]. While clinical data interoperability is the overarching goal, interoperability of the terminology services themselves can be a catalyst to achieving this. The standardization of terminology services can not only provide predictable integration patterns – and thus, greater uptake – but also promote sharing of terminology resources across the enterprise and beyond.

**Quality Attribute 2: Adaptability.** The use cases for terminology systems are continually changing and evolving. Meaningful Use, a large incentive program for leveraging value from the EHR, phases in new functionality over several years[34]. The standards landscape is also changing. Specifically, the increasing use of Fast Healthcare Interoperability Resources (FHIR)[35] will place new demands on terminology systems. Responding to stakeholder needs is critical, and we expect these needs to evolve over time[36]. We use this quality attribute to examine how easily and quickly the system may be changed to meet these evolving needs[32].

**Quality Attribute 3: Usability.** This quality attribute measures the “quality of a user’s experience”[32], with a particular focus on user goals, motivations, and background. Software tooling for terminologies cannot be usable by terminologists alone – it takes a varied mix of subject matter experts to extract the maximum knowledge from the data[37]. Our target user base is highly skilled and often specialized, so impeding these users with usability issues is an expensive proposition. Historically, however, many health care systems are difficult for the intended users to use because they lacked a user-centered design[38]. This leads to poor adoption and can have a direct impact on system success[39]. As usability is of particular interest to our system, we utilize the TURF Framework as a design-time guide – or as the framework states, guidelines for “built-in usability”[40]. This framework, originally designed for analyzing EHR usability, focuses analysis into four components: Task, User, Representation, and Function. Respectively, these components examine the steps or tasks that the user must complete, who uses the system, how the tasks are represented...
to the user, and the basic function the system is to perform. Using this framework will scope our current qualitative design-time discussion and provide a platform for future quantitative usability tests.

**IMPLEMENTATION**

With a concrete use case providing functional requirements and enumerated quality attributes, we now have a mechanism for analyzing an implemented system. Our prototype system implementation is based on three main pillars: (1) a Common Terminology Services 2 (CTS2)[41] compliant network of terminology services, (2) a user-centered graphical user interface to expose functionality to end users, and (3) a loosely-coupled, componentized architectural style. These three implementation focuses are described further in the sections below.

**Terminology Service: CTS2**

CTS2 is an Object Management Group® (OMG) terminology services standard, and includes both a structural model and service specification. It is based on Representational State Transfer (REST)[42] principles, and is comprised of a Platform Independent Model (PIM) and two Platform Specific Models (PSM)s.

**PIM.** CTS2 is specified using Unified Modeling Language® (UML)[43]. Using a platform-independent modeling representation allows for (1) an implementation-independent yet unambiguous specification, and (2) the ability to derive platform-specific models tailored to a specific implementation or computational platform.

**PSM.** The OMG specification provides two platform-specific models, Hypertext Transfer Protocol (HTTP)[44] REST and Simple Object Access Protocol (SOAP)[45]. For our implementation, we will focus primarily on the REST PSM.

**Implementing the Specification.** To speed development, the CTS2 Development Framework was used as a base for implementation. This open-source project is a toolkit for building CTS2 compliant applications, providing compliant HTTP REST bindings, parameter validation, error handling, and a variety of other features. This leaves implementers responsible for more use case specific tasks, such efficient storage and retrieval of terminology data.

To align with the REST architectural style, the CTS2 specification is divided into the structural representation of resource types (Structural Profiles) and a common interface to interact with them (Functional Profiles). As stated above, our system at a minimum must support value set and mapping resources. This requires, at a minimum, a CTS2 service which supports the VALUE_SET_DEFINITION and MAP_VERSION structural profiles. Functionally, to allow full read/write access to these resources, the READ, QUERY, and MAINTENANCE functional profiles should be implemented. See the full OMG specification (http://www.omg.org/spec/CTS2/) for more information on these profiles.

**Graphical User Interface: The CTS2 Workbench**

User experience is of critical importance to our system, and nothing factors into user experience more than the user interface. It is important, however, that although the system is standards-based, it is user-centered. This means that to maximize usability, user interfaces must be in alignment with the cognitive models of the intended users, not simply expose the standardized functionality. For our prototype system, we propose the CTS2 Workbench, a web-based platform for interacting with terminology resources. Although functionally based on CTS2, the Workbench aims to abstract as much detail as possible away from the end user, favoring user task accomplishment over 100% CTS2 functionality exposure. Selected functional capabilities are further described below.

**Value Set Import.** In keeping with our user-centered approach, it is important to understand how value sets (or data dictionaries) are currently being managed. Spreadsheets are ubiquitous in the enterprise, readily available to end users, and for better or worse, serving important roles in business function[46]. For our purposes, they are also a simple storage mechanism for tabular data – a schema in which value sets and mappings may be readily fit. Rather than force an alternative workflow onto the user, we prefer to leverage the existing spreadsheets by viewing them as part

---

2 Source code available at https://github.com/cts2/cts2-framework
of the workflow[47]. Figure 1 shows how the CTS2 Workbench, via a graphical value set import wizard, can accept copy/pasted spreadsheet data and consume it as a CTS2 value set. We assume no a priori knowledge of the imported data structure, so the user is required to apply the appropriate metadata to the parsed columns via a drag-and-drop interface.

![Figure 1: Importing a CTS2 value set using data from a spreadsheet with user-applied drag-and-drop metadata.](image)

**Mapping Creation.** At a high level, terminology mapping is an exercise in linkage. This is important to note, because through past experiences and daily environmental interactions, users may have pre-conceived cognitive models of how linkages are represented. We can capture these notions as image schemas, or representations of these existing models[48]. Constructing a CTS2 MAPENTRY resource is the functional goal of mapping concepts, but approaching this activity with the linkage image schemas in mind will improve usability. Intuitively, a line connecting two objects is a strong way to convey linkage[49]. Figure 2 contrasts an approach based heavily on the CTS2 structure (Detailed View) with one aligned to a linkage image schema (Simple View). The CTS2 Workbench is implemented to allow both views to be used, giving users the option of either a simplified interface or a more powerful, detailed one. This layered approach to interface complexity is used throughout the CTS2 workbench and can be an effective way of engaging users with different levels of experience or domain knowledge[50].

**Architectural Style: Microservices**

Microservices is an architectural style where targeted, task-specific software modules are implemented, deployed, and maintained independently, while together functioning to fulfill a larger task[51, 52]. Using this style can allow for rapid, incremental expansion of the system as needs arise. For example, in our prototype, functionality was added to the system to automatically map two value sets based on lexical similarity. This presented two challenges: (1) it was outside of the scope of the CTS2 specification, and (2) computing lexical similarity is a complex task. By deploying this functionality as a microservice, we are able to wrap CTS2 services without modifying the larger CTS2 infrastructure. Also, because this service has a separate development/deployment lifecycle, we are free to redeploy rapidly as our lexical mapping implementation improves.
Figure 2: Using established cognitive models to abstract CTS2-specific details when creating terminology mappings.

**DISCUSSION**

**Quality Attributes.** Seventy value sets specified by the PCORnet CDM were imported into the system, as well as an increasing number of internal value sets and mappings from various sources. The CTS2 specification demonstrates the ability to model these constructs without degradation of semantics. More importantly, via microservices, our architecture is positioned to respond to evolving user demands. This demonstrates a degree of **Adaptability**.

Leveraging existing terminology sources reduces the system implementation burden. By utilizing the CTS2 specification, we are able to consume data from the National Cancer Institute (NCI) CTS2 implementation. This rich source of data provides numerous standard vocabularies, eliminating the need for us to manage them ourselves. Because the data and service interactions have been standardized, the CTS2 Workbench is able to consume both data points seamlessly. Without the **Interoperability** enabled by CTS2, system cost would have been much higher.

Centering the design on the user drove much of the implementation. At the beginning of the LHSNet effort, all value sets and mappings were managed on spreadsheets. Rather than disregard this process as inefficient, it was leveraged as insight into the cognitive models of end users. Placing a focus on importing data from spreadsheets also allows us to lower the barrier to entry of the system. Focusing on the user, their existing processes, and their cognitive models allows the system to have high potential **Usability**.

**TURF Framework.** The four components of the TURF Framework played a large role in the usability design. First, the **User Analysis** focused mainly on ETL developers engaged in creating the PCORnet CDM compliant data mart for the LHSNet project. This analysis, conducted using semi-weekly meetings, provided an overview of their experience with terminology services as well as their motivations for applying these services to an ETL process.

Next, LHSNet and two related use cases (i2b2 and OHDSI) were examined for overall functional requirements. This allowed us to derive some generalized functionality statements necessary to support creation of the PCORnet CDM compliant data mart, our primary use case. Aligning the design to concrete use cases was part of the **Functional Analysis** of the system.

After analyzing functionality, **Task Analysis** was used to determine the steps currently taken by users to facilitate the ETL process. By examining the users’ current workflow and tools, we were able to align the architecture to incorporate

---

3https://wiki.nci.nih.gov/display/LexEVS/LexEVS+6.2+CTS2
existing spreadsheet-based methods by providing import functionality. Furthermore, through task analysis, we can isolate steps in the user processes that can be streamlined with additional tooling support. Using microservices, we can quickly respond to this need, as demonstrated by the automated mapping example.

Finally, Representation Analysis was used to ensure user interface components expressed these tasks in straightforward ways. Through interviews of LHSNet ETL developers and an analysis of their current tools, alignment of the design to their existing cognitive models could be measured. Through this process, mapping visualization was identified by the users as a deficiency of the spreadsheet-based approach. An image schema was developed for this feature in collaboration with the users and served as the basis for implementation, shown in Figure 2. Iterative development with user feedback was used to validate and refine the implementation.

**Summary.** We find that these quality attributes along with the TURF Framework are effective architectural tools for designing a user-centered terminology system. Through our prototype implementation, we demonstrate that the design-time emphasis on usability, interoperability, and adaptability can be successfully applied to the development of a terminology system, maximizing system utility.

**Limitations and Future Work.** For the LHSNet project, semantic concept to concept alignment was not necessarily sufficient. Smoking Status was one such example. As Figure 3 shows, the alignment of semantics sometimes depends on business logic. In this example, we attempt to align the internal semantics of tobacco use with the PCORnet CDM. To be classified as a PCORnet smoker, we must assert that tobacco is currently in *use* and that the tobacco is being smoked. As shown by the PCORnet CDM value set, further classification of *light/heavy/etc.* is possible, introducing the need for further business rules. Although the CTS2 specification supports a rule-based mapping, it is unspecified as to how those rules should be encoded or executed. Furthermore, user interaction with business rules (such as authoring and testing) is often difficult for non-technical users[53], and would introduce new usability challenges.

Although the TURF Framework served our need as a design-time guide for building-in usability, a quantitative usability study using this framework was beyond the scope of this work. As the prototype implementation matures, the TURF Framework can be used to gather usability metrics (such as task completion time), which in turn can be used to provide evidence-based recommendations for future development iterations.

![Figure 3: Intermixing business logic and semantic mapping: A flowchart for aligning Smoking value sets.](image)

**Acknowledgement.** This study was supported in part by the LHSNet project (CDRN-1501-26638) and the caCDE-QA project (U01 CA180940). The authors also thank Rick Kiefer for his review.
References


electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American
[29] Jiang G, Solbrig HR, Iberon-Hurst D, Kush RD, Chute CG. A collaborative framework for representation and
harmonization of clinical study data elements using semantic MediaWiki. AMIA Summits Transl Sci Proc.
[31] Boehm BW. Verifying and validating software requirements and design specifications. IEEE Software.
[34] Department of Health and Human Services, Office of the National Coordinator for Health Information Tech-
nology. Request for comment regarding the stage 3 definition of Meaningful Use of Electronic Health Records
308.
[38] Johnson CM, Johnson TR, Zhang J. A user-centered framework for redesigning health care interfaces. Journal
for Success. 2015.p. 41.
[42] Fielding RT. Architectural styles and the design of network-based software architectures. University of Califor-
nia, Irvine; 2000.
RFC 2616, June; 1999.
[46] Panko RR, Port DN. End user computing: the dark matter (and dark energy) of corporate IT. In: System Science
[48] Loeffler D, Hess A, Maier A, Hurtienne J, Schmitt H. Developing intuitive user interfaces by integrating users’
mental models into requirements engineering. In: Proceedings of the 27th International BCS Human Computer
Can Patient Record Summarization Support Quality Metric Abstraction?

Rimma Pivovarov, PhD1, Yael Judith Coppleson, MPH1, Sharon Lipsky Gorman, MS2, David K. Vawdrey, PhD1,2, Noémie Elhadad, PhD2

1Value Institute, NewYork-Presbyterian Hospital, New York, NY; 2Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

We present a pre/post intervention study, where HARVEST, a general-purpose patient record summarization tool, was introduced to ten data abstraction specialists. The specialists are responsible for reviewing hundreds of patient charts each month and reporting disease-specific quality metrics to a variety of online registries and databases. We qualitatively and quantitatively investigated whether HARVEST improved the process of quality metric abstraction. Study instruments included pre/post questionnaires and log analyses of the specialists’ actions in the electronic health record (EHR). The specialists reported favorable impressions of HARVEST and suggested that it was most useful when abstracting metrics from patients with long hospitalizations and for metrics that were not consistently captured in a structured manner in the EHR. A statistically significant reduction in time spent per chart before and after use of HARVEST was observed for 50% of the specialists and 90% of the specialists continue to use HARVEST after the study period.

Introduction

In 1999, the Institute of Medicine released To Err is Human,1 which elucidated shortcomings in healthcare and ignited the movement to improve quality and patient safety. Many initiatives aimed at improving healthcare quality were developed, several of which involved the creation of large-scale data registries that are populated using both claims data and manual data abstraction.2

Indicators used to assess quality of care are often buried within patient records. To accurately abstract these quality indicators, specially trained nurses manually comb through patient records to locate relevant information. Our 2,600-bed institution employs 35 full-time data abstraction specialists dedicated to reporting quality metrics for 30 databases covering 13 disease states and processes of care. Measures include CMS and Joint Commission databases for Core Measures, as well as disease-specific registries such as transplant, sepsis and stroke. The goals of the databases vary, from national benchmarking, to inclusion in value-based purchasing, to peer-comparisons for facilitating quality improvement. Participation in some databases is voluntary, such as the Society for Thoracic Surgery data registry, while participation in others is a regulatory requirement, such as the UNOS transplant database.

Each data abstracter is responsible for extracting a particular set of data using a combination of structured and unstructured clinical data including laboratory test information, comorbidities and complications through review of clinical documentation, and changes in clinical status throughout a patient’s hospital stay. The complexity of each database and ease of access to information is highly variable. Some registries populate data through electronic feeds of structured documentation while others require complete manual abstraction for everything from demographics to time of symptom onset. Although every data abstracter is responsible for reporting information from patients’ heterogeneous and voluminous records, the information they seek is very different, and the tools they use are different as well.

While there are many benefits in participating in quality improvement registries and databases, the burden of manual abstraction can be excessive. In fact, it is estimated that for physician practices, physicians and staff spend on average 15.1 hours per week and approximately 15.4 billion dollars annually dealing with the reporting of quality measures.7 Despite the large time burden, it is likely that the number of reported quality metrics will continue to grow.8

There has been some recent success in developing natural language processing (NLP) approaches for quality metric abstraction. However, these approaches are evaluated in a retrospective fashion, outside the workflow of clinical data abstraction experts, and have largely been disease- or metric-specific. Some tools have been developed mainly for internal comparison purposes3 and others optimized to report quality metrics to specific databases: Yetisgen et al. developed a system for automated extraction of comorbidities, operation type, and operation indication for surgical care reporting4 and Raju et al. developed an automated method for the specific extraction of adenomas mentions5.
Beyond the barriers of moving from retrospective studies to interventions, and deploying NLP-based systems in a hospital setting, there exist barriers specific to the DA work: the metrics, diseases, and types of information needed vary drastically from one database to another. Furthermore, most of the DA work consists of abstracting (rather than extracting specific facts from the record) based on unstructured guidelines (although there has been work on automatically structuring quality metrics\(^6\)). The disease-specific work has yielded promising results and shown that some disease-specific metrics can be extracted in a fully automated fashion. Still, to date there are few available disease-specific quality metric abstraction tools as they are difficult to create and optimize, especially as the gathering of quality metrics is very heterogeneous. We hypothesize that a more holistic and broad approach could provide much needed support to a variety of DAs at once.

There has been some research demonstrating the potential benefit of holistic clinical summarization tools for physicians at the point of care\(^9,10\), however, there has been much less work examining the information needs and utility of a holistic summarization tool for the purpose of quality abstraction. Our institution already provides EHR users access to HARVEST\(^11\), a real-time, patient-level, summarization and visualization system. HARVEST was originally developed to aggregate relevant patient information for busy clinicians who often do not have the time to read or even skim through all available previous clinical notes. In this study, we explore the value of HARVEST in a different scenario—as a support tool for DAs in their abstraction work. HARVEST’s interface consists of three sections: a Timeline, a Problem Cloud, and a Note Panel (Figure 1). The Timeline is interactive and provides a high-level display of a patient’s inpatient, outpatient, and ED visits. The Problem Cloud displays all of the patient’s documented problems (as extracted from parsing of all the patient’s notes) with the most salient problems for this patient during the selected time range appearing larger and on top. HARVEST is a general-purpose summarization system: it extracts all problems relevant to a patient and operates on all patients in the institution. As such, it differs from the dedicated NLP approaches to quality abstraction described above in that it does not aim to do the work of the DAs, but rather support them in their search for information in the vast amount of documentation available for a given patient.

![Figure 1](https://example.com/harvest-screenshot.jpg)

**Figure 1.** Screenshot of HARVEST for a de-identified patient record. The Timeline focuses on a single admission, and the user has selected "vomiting" in the Problem Cloud. The Note Panel shows the notes for that visit which mention the problem. The user has selected specific note (on the right) and all the mentions of that problem are highlighted.
Our overall research question for this study is whether a patient record summarization tool, such as HARVEST, supports the needs of DAs during their abstraction work. Towards that goal we designed a pre/post intervention study aimed at DAs with a heterogeneous range of metrics and abstraction workflows. We investigated the DAs’ subjective satisfaction of the tool as well as the impact on workflow based on their usage logs of the EHR.

Methods

Data Collection

After obtaining appropriate institutional review approval, we presented HARVEST in November 2015 to Data Abstraction specialists at NYP as part of one of their “Lunch and Learn” meetings. The presentation included a 10-minute demonstration of HARVEST and where to access it on the EHR. DAs were recruited to participate in our study and consented one by one.

DA study participation consisted of four parts: (1) subjects were asked to answer a pre-study questionnaire; (2) for 4 weeks, the subjects were asked to use HARVEST during their normal course of abstraction, whenever they felt it might be useful to their abstraction and for any given patient; and (3) after the study period, the subjects were asked to respond to a post-study questionnaire. Finally, for all subjects, (4) their EHR usage logs were collected for three months before the beginning of the study and two months after the end of the study period (i.e., three months after beginning of study).

The pre-study questionnaire consisted of 20 questions and captured basic information about the DAs, including how long they had done abstraction work, which database they abstract for, how many data elements they need to abstract for a given patient chart, and a description of their typical workflow when abstracting a chart (i.e., which part of the EHR they visit and in which order). The pre-study questionnaire also asked for expectations of where HARVEST might be most useful, and where it may fit into the DA’s specific workflow.

During the 4-week study period, we sent weekly reminders to the subjects that they could use HARVEST whenever they felt it would be useful to them.

The post-study questionnaire consisted of 33 questions. The questions were inspired by the Technology Acceptance Model (TAM) and the System Usability Scale (SUS). Many of the SUS questions were included in the questionnaire to measure ease of use and intention to use. Other questions were included to capture the DA’s overall perception of HARVEST, for what purposes they used HARVEST, its perceived usefulness for their abstraction work, whether they had identified any unintended consequences of using HARVEST, and whether the subjects planned to continue using HARVEST in their daily work.

The questionnaires were distributed and results were collected using Qualtrics. The EHR usage logs were obtained for all systems in the NYP EHR ecosystem (which include HARVEST usage logs). The logs contained information about which function within the EHR was accessed and at what time. Examples of action types include document review, laboratory test results overview, and visit summary review. Usage log analysis enabled the computation of metrics such as time spent abstracting a given patient chart and how many patients were reviewed with or without access to HARVEST. Basic workflow information can also be derived from the usage logs as frequent sequences of actions of the different users.

Metrics

We assessed the impact of HARVEST on DA workflow in three different ways: user perception, workflow changes, and user retention. All data processing and statistical analyses were done using a combination of python and R.

User perception about HARVEST was measured through the pre and post-questionnaire questions. We compare how the DAs predicted they would use HARVEST and how they actually used HARVEST during the study period.

For the workflow analysis, we present DA self-reported workflows and measure workflow changes using the EHR log data. Workflow changes were measured in two different ways: how long DAs spend and how many EHR action types DAs take on individual patient abstractions before and after the introduction of HARVEST. The pre-HARVEST period was defined as 3 months before the DA was introduced to HARVEST, and the post-HARVEST period was defined as 3 months after the DA was introduced to HARVEST. A t-test with Bonferroni correction was used to measure whether there was a change in time spent per patient or a change in number of EHR actions for each DA before and after the introduction of HARVEST. Of note, all patients abstracted in the post-HARVEST period (whether seen with HARVEST or not) were included in the EHR log analyses.
To measure the most common workflows within HARVEST, we ran the SPADE algorithm\textsuperscript{1} for sequence event mining (using the cSPADE package in R) for all HARVEST accesses during the study period.

Retention was measured using EHR logs to see how many DAs returned to HARVEST after the study was completed and how often they used HARVEST after the study was completed. The log-based retention was compared to how the DA predicted they would use HARVEST in the future.

**Results**

**Subjects and DA abstraction tasks**

10 DAs out of the 35 DAs enrolled in the study. Subjects were nurses, and most have worked as DAs for less than 5 years (minimum experience was 1 year, maximum was 13 years). Each completed both the pre-questionnaire and the post-questionnaire 4 weeks after. The median time to completion for the pre and post-study questionnaire was 25 and 13 minutes, respectively.

Together, the enrolled subjects abstract quality metrics for 7 different databases, corresponding to 6 diseases of interest (cardiac metrics are reported to multiple databases). Databases to abstract patient information varied in the number of data elements to abstract, as well as the number of forms to fill out per patient: some have 6-10 different forms per patient while others only 1. In addition, DAs may return to the same patient over time; for example, transplant DAs fill out one form for a patient pre-transplant and another for the same patient at 6 months post-transplant.

Table 1 summarizes the self-reported DAs experiences and workflows, as captured by the pre-study questionnaire. A majority of the DAs abstract over 75 individual data elements for each patient. The data elements are found in at least 7 different EHR data types (e.g., physician and nurse notes, laboratory tests, flowsheets, medication orders, etc.). To find each of the 75+ data elements the DAs routinely visit multiple clinical systems during their workflow, likely a consequence of the fragmented and legacy systems that house patient data in our large academic medical center. One commonality across all DAs is they reported that most of their time is spent reading and abstracting from clinical notes.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Database</th>
<th># DAs who abstract for disease enrolled in study</th>
<th>% DAs who abstract for disease who enrolled in study</th>
<th>Average # of data elements abstracted per patient</th>
<th># different EHR data types accessed during abstraction</th>
<th># different systems accessed during abstraction</th>
<th>Where most of DA's time is spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bariatric</td>
<td>Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program</td>
<td>1</td>
<td>100%</td>
<td>100+</td>
<td>7</td>
<td>3</td>
<td>Notes</td>
</tr>
<tr>
<td>Cardiac</td>
<td>Society of Thoracic Surgeons NY State DOH</td>
<td>3</td>
<td>29%</td>
<td>100+</td>
<td>8</td>
<td>2-4</td>
<td>Notes, Operative Data</td>
</tr>
<tr>
<td>Sepsis</td>
<td>NY State DOH</td>
<td>2</td>
<td>57%</td>
<td>75-100</td>
<td>7</td>
<td>2-4</td>
<td>Notes, Vital Signs</td>
</tr>
<tr>
<td>Stroke</td>
<td>Get With the Guidelines</td>
<td>1</td>
<td>25%</td>
<td>75-100</td>
<td>7</td>
<td>1</td>
<td>Notes</td>
</tr>
<tr>
<td>Surgery</td>
<td>American College of Surgeons</td>
<td>1</td>
<td>50%</td>
<td>100+</td>
<td>7</td>
<td>2</td>
<td>Notes</td>
</tr>
<tr>
<td>Transplant</td>
<td>United Network for Organ Sharing</td>
<td>2</td>
<td>35%</td>
<td>25-50</td>
<td>10</td>
<td>3-4</td>
<td>Notes, Labs, Meds</td>
</tr>
</tbody>
</table>

Table 1. Participant-reported data from the pre-study questionnaires on what is abstracted, from where, and what is the DA’s general workflow.

When asked further about their abstraction workflow, 9 out of 10 subjects reported that finding relevant information from clinical documentation was tedious, and only 30% of DAs agreed that patient chart review was efficient. There
was in fact a very wide range of perceived time spent per patient chart review across DAs and in particular across databases. Stroke DAs reported spending on average between 20 and 45 minutes per patient; bariatric reported 35-50 minutes per patient; surgery reported 45 minutes per patient; transplant DAs reported between 20 and 90 minutes per patient; cardiac DAs reported 20-45 minutes per patient for one database and 30-120 minutes for another; and sepsis DAs reported 120-180 minutes per patient on average.

**DAs Satisfaction with HARVEST**

In the post-study questionnaire, all subjects agreed that HARVEST was accurate, all elements in the HARVEST interface were necessary and not redundant. All subjects felt confident using HARVEST and did not perceive any negative impact when using it for patient chart review. In addition, most subjects (80%) found the system easy to learn, easy to use, and not unnecessarily complex.

The majority of subjects thought that HARVEST positively impacted their abstraction process (60%), while the rest saw no impact on their abstraction process (40%). None saw a negative impact. Similarly, 60% of the subjects thought they would continue using HARVEST. When asked when HARVEST was useful and when it was not, subjects provided feedback as summarized in Table 2. Subjects found HARVEST most useful when patients had longer charts and they liked the problem-oriented approach to navigating the record. Subjects who did not find HARVEST helpful to their abstraction process reported that the data they abstract is typically easy to find in the patient chart already.

| HARVEST is most useful in when patients have longer charts | “It was helpful in all the cases I abstracted, but definitely helped me save time with longer charts, as we have many elements to abstract. It helped me narrow down timing of events such as onset of sepsis or comorbidities/complications during the patient’s stay” |
| No added value to HARVEST when the information needed is easy to access in the chart. | “Most of the patients I’m abstracting have very short visits so it's easy to read the chart through. Most of the data points are related to the current visit so it's easy to get all the information directly from the short charts. The only data points that Possibly helpful was the history. If this wasn’t included in the H&P (which is rare) then I would try to use HARVEST for it.” |
| HARVEST’s problem-oriented navigation of the chart is helpful. | “It’s easier to find the notes with correlating information” |

Table 2. Common perceptions of HARVEST from subjects and quotes from the post-study questionnaire.
Some DAs found that HARVEST played a positive role in data verification (2 found it to be very helpful, and 2 sometimes used it for verification). 6 DAs found that HARVEST was able to help identify things that they would have missed while others thought they would have found the information, but perhaps it would have taken longer.

**Workflow**

While the DAs were able to explain their workflow (i.e., series of actions in the EHR to carry out the abstraction for a given patient) in a detailed fashion in the pre-study questionnaire, the workflows were not as easily discernable when analyzing the EHR usage logs. Nevertheless, there were some clear patterns identified when comparing EHR usage logs 3-months pre-study, and 3-months after introduction to HARVEST.

**Time spent on patient abstraction.** When looking at the distribution of time spent per patient before and after HARVEST, and binning the distributions into quartiles, we found statistically significant reduction in time spent on patient abstraction for 50% of the subjects: one subject had a reduction in overall time spent over all quartiles resulting in an average 20 minutes time gained per patient, two subjects saw a reduction in time spent on short abstractions (first quartile of the distribution), and two subjects saw a reduction in time spent on long abstractions (last quartile of the distribution). For illustrative purposes, Figure 2 displays examples of distributions time before and after the introduction of HARVEST. Additionally, 1 DA saw a statistically significant increase in time spent post-HARVEST.

These findings from the EHR usage logs were confirmed by the subjects’ perceived impact of HARVEST on their workflow: 4 of the 5 DAs who had statistically significant shorter times reported also reported that their workflow was shortened by HARVEST. The DA who had an increase in time spent on abstraction post study did not report a perceived increase.

![Figure 2](image-url) Density plots of the time spent per patient pre- and post-HARVEST introduction. A t-test demonstrated that the sepsis abstractor had a significant reduction (p < 0.001) in total time spent. On average, the sepsis abstractor saved 20 minutes per patient. The bariatric DA had no statistically significant reduction in time spent on chart review.

**EHR actions.** There was a significant decrease in EHR actions related to accessing the list of visits and information related to the visits. In fact, the only EHR action that increased pre- and post-HARVEST introduction was access to the HARVEST tab in the EHR. Table 3 displays some of the statistically significant differences in EHR actions and their usage pre- and post-HARVEST introduction. We hypothesize that the decrease indicates that subjects used the Timeline view of HARVEST, which visualizes all of the patients’ visits from all settings, as an alternative to the traditional EHR visit list. EHR access to notes was also decreased for 3 DAs. This finding was also confirmed by the
perceived impact of HARVEST as reported in the post-study questionnaire, where access to documentation through the problem-oriented view was seen as helpful to the abstraction process.

<table>
<thead>
<tr>
<th>Average Actions per Patient</th>
<th>Pre HARVEST</th>
<th>Post HARVEST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Review of Documentation</strong></td>
<td>95.72</td>
<td>63.14</td>
</tr>
<tr>
<td></td>
<td>58.23</td>
<td>45.43</td>
</tr>
<tr>
<td></td>
<td>14.57</td>
<td>11.75</td>
</tr>
<tr>
<td><strong>Check Orders</strong></td>
<td>17.17</td>
<td>7.03</td>
</tr>
<tr>
<td></td>
<td>1.63</td>
<td>1.19</td>
</tr>
<tr>
<td><strong>View Patient</strong></td>
<td>7.76</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>4.48</td>
<td>3.63</td>
</tr>
<tr>
<td><strong>Access Patient Demographics</strong></td>
<td>3.13</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td>3.64</td>
<td>2.84</td>
</tr>
<tr>
<td><strong>Visit Summary</strong></td>
<td>2.36</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>2.68</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Table 3. All EHR actions with a statistically significant pre- and post-HARVEST difference for at least two DAs. Each of the rows displayed has a Bonferroni corrected p-value < 0.001. Documentation review had a significant decrease for 3 DAs.

**HARVEST workflow for patient chart abstraction.** A deeper dive into the workflows within HARVEST found that aside from the basic action of loading HARVEST and setting a time period to view, the most frequent parts of HARVEST used are the selection of a problem in the Problem Cloud and opening of a note. Across the entire study period and 10 DAs, there were 222 sequences where HARVEST was used. Table 4 shows the 11 possible actions as part of the HARVEST interface, and the usage frequency of each by the subjects during the 3-month post-HARVEST introduction period.

<table>
<thead>
<tr>
<th>% of HARVEST workflows that incorporate this action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access HARVEST tab in EHR for a patient</td>
</tr>
<tr>
<td>Set Timeline to a specific date range</td>
</tr>
<tr>
<td>Select problem in Problem Cloud</td>
</tr>
<tr>
<td>Open Note</td>
</tr>
<tr>
<td>Expand Problem Cloud to include all problems</td>
</tr>
<tr>
<td>Zoom in on Timeline</td>
</tr>
<tr>
<td>Select a particular visit/admission in Timeline</td>
</tr>
<tr>
<td>De-select a problem in Problem Cloud</td>
</tr>
<tr>
<td>De-select a particular visit/admission in Timeline</td>
</tr>
<tr>
<td>Shrink Problem Cloud to most salient problems</td>
</tr>
<tr>
<td>Zoom out to full longitudinal view of patient chart</td>
</tr>
</tbody>
</table>

Table 4: The usage frequency of each of the 11 actions across all DAs in the study.

As complement to these findings from the log analysis, the pre- and post-study questionnaire showed the following trends: 80% of the subjects expected the Timeline to be useful pre-study, but only 60% found it useful after the 1-month study period. 70% of the subjects reported finding the note access functionality in HARVEST useful, which is
also reflected in the 72% usage of the “Open Note” functionality based on the EHR logs. And similarly, 60% of the subjects found the Problem Cloud useful for their abstraction workflow in the post-study questionnaire; also reflected by the 68% usage of the “Expand Problem Cloud” function in HARVEST (Table 4).

Figure 3 shows the most common sequences of HARVEST actions, as derived from all subjects in the 3-months post-HARVEST introduction. Most commonly DAs just went straight to note without clicking term. Many times they would select and re-select the timeline, move the timeline (Figure 3). Multiple “Load HARVEST for Patient” within one sequence occur if the DA accesses another tab and returns to HARVEST or refreshes the HARVEST page.

**Figure 3**: Most frequent workflows in the HARVEST application during one session identified using the SPADE algorithm. The width of the arrow represents frequency with which this event sequence occurs.

**HARVEST Usage and Post-Study Retention**

Overall, 70% of the study participants used HARVEST for at least 10% of their abstractions. The 2 Sepsis DAs used HARVEST on over 80% of their patient abstractions.

When the DAs were asked if they would continue to use HARVEST, 60% predicted that they would. Most commonly these DAs found that HARVEST would continue to be a part of their work because it helps filter and search through notes in a fast and accurate way, it has the ability to enable search for conditions and events that would be hard especially during extended admission, and it saves time. For the 40% who predicted they would not use HARVEST in the future, the problem-oriented view of the record presented by HARVEST did not align with the type of information they abstract in their tasks (for instance, one DA said “None of the problem cloud words are ones I would use.”), or that the data they needed was already easily accessible in the chart (see quotes in Table 2).

Interestingly, even though only 60% of subjects said they would continue using HARVEST, 90% of the study participants continued to use it in the post-study period of two months (see Figure 4). Even for some of the subjects for whom there was no significant time gained in patient chart abstraction, they continued using the tool for a significant portion of their charts (e.g., bariatric DA). The main exception to post-study retention were the cardiac DAs: 2 of the 3 abstractors predicted that they would continue to use HARVEST; however, according to the EHR logs their usage of the tool dropped after study completion. It is not clear exactly why the cardiac abstractors used HARVEST less than expected, but it seems that the tool did not become well integrated into their workflow. Alternatively, the sepsis team, which has one of the longest and most complex abstraction workflow (up to 180 minutes pre-HARVEST), saw such a change in workflow that they adopted the use of HARVEST as part of their protocol for chart abstraction.
Figure 4. The retention rates of the DAs use of HARVEST. HARVEST usage stayed fairly steady after the study period for all metric types except cardiac where the use was much reduced, and stroke where the use more than doubled.

Discussion

The pre-study questionnaire reflected the complex and widely diverse activities involved in the process of chart abstraction across diseases of interest and quality databases.

Questionnaires and EHR usage log analyses provided complementary views of the impact of HARVEST on abstraction tasks across diseases. Our findings suggest that a problem-oriented patient summarizer coupled with a patient timeline can support the nurses and their information needs. HARVEST is most useful for DAs who are asked to locate data elements for patients with lengthy hospitalizations (and thus with large amounts of clinical documentation in their charts) as well as to locate data elements that are distributed in many parts of the record. However, the extent to which a summarizer such as HARVEST can be useful is variable and depends on the specific abstraction task, as well as to which extent the patient chart already documents the desired data elements in one place.

In general, DAs had a mostly positive response to HARVEST and found it useful for data verification, searching within clinical notes, and identifying the timing of events. We also found that HARVEST was able to provide statistically significant time savings for some groups of DAs. Although not all of the DAs had a statistically significant reduction in the time spent abstracting and some reported that HARVEST did not have specific terms that were necessary for their specific abstraction, 90% continued to use HARVEST after the study completion of the study.

In addition to the 10 staff who participated in the HARVEST study, the use of HARVEST continues to spread across the Division of Quality and Patient Safety at NYP. Sepsis abstractors have recently trained Trauma abstractors on how Sepsis uses HARVEST and where it fits best into their workflow. 3 DAs abstracting for trauma, cardiac, and core measures who were not enrolled in the study have been found to consistently using HARVEST. Finally, 1 DA who was recently reassigned to a new database has switched from never using HARVEST to visiting the HARVEST for 56% of their patient abstractions.

Regarding our initial question on the usefulness of a general-purpose summarizer for abstraction, our results indicate that HARVEST supports the needs of abstraction for up to 10 out of the 30 quality databases at NYP.

Limitations

Even though the DAs workflows are well-defined and are without any colleague or patient interruptions like for clinicians at the point of care\textsuperscript{15}, the workflows of the DAs are still complex and unique to each. As such, the EHR usage log analysis was not successful at generalizing workflows similar to the DA self-described ones. Finally, this study is limited to one academic medical center, where HARVEST is deployed. However, it is clear that the quality abstraction burden is ubiquitous across the nation\textsuperscript{7}.

Future Work

Following the most common suggestion from the DAs, we are working to incorporate more data types into HARVEST, presenting salient problems in clinical notes alongside laboratory test results, medications, and billing codes. This is
a non-trivial task, as we want to present all these data types in an accurate and coherent fashion, all the while keeping the problem-oriented view of the record.

It has been suggested that HARVEST could support not only researchers conducting chart reviews, primary care clinicians at the point of care, and emergency department physicians working to obtain a brief overview of a patient they had never seen before, but also EHR users involved in other initiatives across the hospital. We hope to investigate whether HARVEST can be a useful tool when conducting root cause analyses for investigation of serious adverse events in the hospital, coding for patient billing, or chart reviews conducted by the legal department.

**Conclusion**

Patient chart abstraction for the sake of identifying quality metrics is a complex and time-consuming activity, yet a critical one for hospitals. Because the reporting of quality metrics requires interpretation of patient data, it is not clear that simple extraction of facts form the patient record (in particular extraction using natural language processing) is the most helpful and robust way to support data abstraction specialists. A general-purpose summarization system, such as HARVEST, which supports a broad array of data abstractors in their abstraction tasks by helping them access salient part of the record, is a promising alternative solution.

**Acknowledgments:** This work was supported in part by a National Science Foundation award (#1344668).

**References**

Ari H Pollack, MD1,2, Andrew Miller, PhD2, Sonali R. Mishra, MS2, Wanda Pratt, PhD2.
1Seattle Children’s Hospital, Seattle, WA; 2University of Washington, Seattle, WA.

Abstract

Participatory design, a method by which system users and stakeholders meaningfully contribute to the development of a new process or technology, has great potential to revolutionize healthcare technology, yet has seen limited adoption. We conducted a design session with eleven physicians working to create a novel clinical information tool utilizing participatory design methods. During the two-hour session, the physicians quickly engaged in the process and generated a large quantity of information, informing the design of a future tool. By utilizing facilitators experienced in design methodology, with detailed domain expertise, and well integrated into the healthcare organization, the participatory design session engaged a group of users who are often disenfranchised with existing processes as well as health information technology in general. We provide insight into why participatory design works with clinicians and provide guiding principles for how to implement these methods in healthcare organizations interested in advancing health information technology.

Introduction

Over the past 20 years, and accelerated by the introduction of the Health Information Technology for Economic and Clinical Health (HITECH) Act1 in 2009, healthcare organizations in the United States have been rapidly adopting health information technology to improve the delivery of health care and reduce medical spending. However, implementing and deploying successful informatics solutions in clinical settings has proved challenging.2 Managers and other C-suite hospital executives often have an incomplete knowledge of the detailed operational workflows the system must support. Thus the primary decision makers often make design choices that are inefficient for end users. Researchers in the field of Human-Computer Interaction have long advocated participatory approaches to system design, in which end users play an active role in the design process, and user-centered design occurs from the beginning of the design cycle. Traditionally, user input or usability testing happens at the end of system development, or at least after major requirements and feature decisions have been set. Health information technology has tended to follow a similar process.3 But within the human-computer interaction community, methods that involve users at earlier stages of the design process are increasingly popular: the earlier users are involved, the more likely it is that the end product will meet their needs. In one such method, Participatory Design (PD), the eventual users of a system contribute to its design in even the most formative stages.4 PD was first developed in Scandinavia in the 1970s as a reaction to the disruptive introduction of computers into the workplace. Those people most affected by information technology—the workers whose jobs were being transformed or even eliminated—had little input into how their information systems were designed, and little power to enact change.5 Participatory design emerged from this context as a set of methods and practices that allow groups of non-designers to engage in discussions around information technology, to develop and refine prototypes, and to express their values with respect to the technology that affects their lives.

Can such approaches work within an existing clinical context? Further: can a hospital involve its own clinicians in designing health related technology? PD has been successfully employed in a wide variety of populations and topics, from marginalized teens6,7 to civic engagement8, even the delivery of healthcare itself9, and is a commonly used technique in Human-Computer Interaction research. PD has also shown promise in the design of healthcare technology, where it has been most frequently employed in the design of patient-facing technology.10 However, given PD’s origins in workplace systems, surprisingly few researchers have used PD to inform the design of clinician-focused systems. Gennari and Reddy introduced the technique to the American Medical Informatics Association community in 2000 in a paper describing the design of a clinical trial protocol management system; however, their paper focused more on interviews and contextual inquiry than on PD workshops as a technique.11 More recently, Kusunoki and Sarcevic, functioning in the role of outside consultants, used PD to inform the design of emergency room awareness systems, including the PICTIVE technique we discuss in this paper.12 Despite the success of these studies, PD work with clinicians has yet to take hold within the Medical Informatics community and little is known about how PD translates into measurable success for health information technology.
In this study, we asked physicians to participate in the design of a novel clinical prioritization tool utilizing PD methods. Our goal was to explore how physicians engage in the PD process and demonstrate their effectiveness in the design of clinical information tools. In previous work, we identified and modified an existing framework, knowledge crystallization, to frame and understand how physicians collect, process, and utilize data during the clinical prioritization process (i.e., identifying individual patients that need attention first). The framework demonstrated that clinicians collect data to categorize and prioritize patients according to an expected clinical course. However, when data does not support their expectations, or when categorization indicates potential for morbidity, physicians increase efforts to act or re-categorize patients. Unexpected clinical changes have a significant impact on the decision-making and prioritization by clinicians. With this foundation, we asked physicians to design a tool to help them with this specific task. In this paper, we identify the factors leading to a successful PD session with physician participants, describe challenges that we experienced, and ultimately provide a series of guiding principles for future physician-focused PD work.

Methods

We conducted an IRB-approved, two-hour participatory design workshop at Seattle Children’s Hospital, located in Seattle, Washington in early 2015. Participants worked individually and in groups to answer the motivating question “Which patient should I see next?”

Recruitment
To be included in this study, we required that physicians be credentialed at either the attending (supervising physician) or fellow (sub-specialty trainee) level, and that they provided acute clinical services to hospitalized patients. The primary author (AP) sent individual emails to 23 different providers at the study site, inviting their participation. We targeted a diverse audience with regards to gender, medical specialty, academic rank, and clinical experience. Given the professional relationships between AP and the medical staff, we identified individuals who we felt would make a positive and significant contribution to our PD session.

Our roles
Three researchers attended the workshop. The first author contacted participants, gave the opening presentation, guided the brainstorming session, and supervised the design activities, asking questions and offering guidance as needed. The other two researchers took notes and photos, documenting groups’ progress and social dynamics. The entire research team provided input on the format and organization of the session.

Session design
We utilized the PICTIVE participatory design technique to provide an environment that supports equal opportunities for a diverse set of participants to engage in the design process. First developed in 1991, the PICTIVE technique was intended to allow those with limited design experience and expertise to have “equal opportunity to contribute their ideas.” Sessions begin with brainstorming activities, designed to stimulate thought and create a dialog between participants, then followed by activities that allow the ideas to be physically expressed through the use of low-tech objects to support the creative process. This process ensures that all are able to contribute their ideas in a meaningful way. While PICTIVE explores the needs of users at a more detailed level, it doesn’t focus on workflow or other system level processes that may support the specific task at hand. Other PD methods, such as CARD, were developed to fill this gap, with the intent to help analyze and design workflows used in software systems. Given our previous work detailed the workflows of physicians for our specific task we asked physicians to focus on, PICTIVE was selected over other participatory design techniques such as CARD given the former’s emphasis on participant generated designs and less on collaborative analysis and workflow.

Materials
We provided a set of craft supplies to aid participants in creating their designs. Participants were able to make use of colored paper, stickers, pom-poms, pipe cleaners, glitter glue, scissors, glue guns, markers, and colored pencils in their designs. Each group was also provided with a foam core board on which they could mount their final design for presentation.

Design workshop overview
We conducted the session in the early evening to maximize participant availability, and provided a boxed dinner for each participant. Before the workshop, we divided participants into three groups: two groups of four, and one group of three. In selecting the groups, we looked for a diversity of experiences and seniority levels. For example, we tried to avoid creating groups where one person was in a group with his or her supervisor. We used participants’ boxed
dinners as place cards; as participants arrived each one sat at their group’s table and socialized before we began the main design activity.

The workshop began with a PowerPoint presentation. The lead author introduced the study, gave an overview of our previous work, and presented participants with the motivating question: “What patient should I see next?” Participants then began the individual brainstorming portion. We asked participants to write down, on notecards, what information they would need, when they would want it, why they would want it, and where they would want it. Participants worked individually for approximately 10 minutes, and then participated in a group brainstorming discussion for the next 20 minutes. We recorded participants’ suggestions on a whiteboard so they could refer to their ideas during the hands on design portion of the workshop.

We then introduced the design challenge and constraints (Table 1). We encouraged participants to use the craft materials and allowed them to divide their time as they wished. Periodically during the design phase, the first author discussed groups’ thoughts and progress with them. In this design review, we sought not to shape participants’ designs but to elucidate their thought processes and stimulate intra-group discussion. In the last 20 minutes of the workshop, participating groups presented their designs to each other.

Data
We collected various types of data in this exercise. We audio-recorded the entire workshop, using smartphones running voice recorder apps at each group table. We also video-recorded the final presentations. One researcher took detailed ethnographic field notes of the entire session, while another researcher alternated between taking notes and taking photographs of groups’ progress. We collected and scanned individual brainstorm cards, and photographed the whiteboard with the brainstorm ideas on it. We collected the artifacts used in the presentations. Finally, the first author had a number of informal conversations with participants in the days following the workshop.

Analysis
Immediately following the workshop, the researchers held a debrief session amongst themselves. We discussed our perceptions of the workshop and proposed initial themes and observations to each other. We listened to the audio recordings and produced a more fine-grained analysis of each group’s work process. We looked particularly for power dynamics within the groups and key ideation points where groups made decisions or shifted phases. We also combined our field notes, notes from the audio-recordings, and photographs of the design sessions to generate a narrative about each group’s progress. Several days after the workshop, we met to share these analyses and refine themes. Finally, we explored each design through a standardized coding tool that had six prompts: (1) What does the design look like, (2) What info is included, prioritized, seen in other designs?, (3) What does the tool do?, (4) What information is missing?, (5) Any new ideas?, (6) How does the design tie back to the original framework?

Results

Participants
We recruited 11 physicians (five females) representing six different medical specialties (Table 2). Five of the subjects had participated in our previous study and therefore had some previous exposure to the topic. We divided the subjects into three groups, two groups of four and one group of three and they sat with their respective groups during all segments of the design session.

Table 1. Design challenge constraints

<table>
<thead>
<tr>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The tools must highlight the information components used in the process of clinical prioritization.</td>
</tr>
<tr>
<td>2. The tool needs to be able to compare parameters across a variety of patients.</td>
</tr>
<tr>
<td>3. The tool should provide enough information to answer the question without the need to open individual patient charts.</td>
</tr>
</tbody>
</table>

Table 2. Participant and group characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants (Female)</td>
<td>3 (1)</td>
<td>4 (2)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Medical Specialties</td>
<td>Craniofacial</td>
<td>Clinical Informatics</td>
<td>General Medicine (2)</td>
</tr>
<tr>
<td></td>
<td>General Pediatrics</td>
<td>General Medicine</td>
<td>Infectious Disease</td>
</tr>
<tr>
<td></td>
<td>Infectious Disease</td>
<td>Nephrology</td>
<td>Nephrology</td>
</tr>
<tr>
<td>Academic Rank</td>
<td>Assistant Professor</td>
<td>Fellow</td>
<td>Assistant Professor (2)</td>
</tr>
<tr>
<td></td>
<td>Associate Professor (2)</td>
<td>Clinical Instructor</td>
<td>Professor (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overall Session Output
During the individual brainstorming session (approximately 20 minutes), ten subjects (one subject arrived late and did not participate in this phase of the session) generated 47 ideas representing approximately 20 different topics or concepts. Individual subjects generated on average 4.7 ideas each (range 3 – 7). Some subjects needed additional clarification and an example to begin the exercise while others had no trouble with the original instructions. When asked to share their individual ideas the most senior physicians spoke up first with the younger physicians contributing later during the discussion, though eventually all subjects participated and spoke up during the group brainstorming session (approximately 20 minutes). This subsequent discussion covered 20 topics or concepts of which 10 were new and not documented during the individual process, suggesting that the group discussion supported the generation of new, more complex ideas (Figure 1).

The sketching and prototyping session lasted for approximately 50 minutes. During this time, groups one and two immediately began working collaboratively generating ideas, while group three started working individually and eventually worked as a group. Despite the productivity and diversity of ideas generated in the brainstorming segment, the prototyping and sketching segment generated far fewer concepts with groups one and two only coming up with one design concept and group three generating four different ideas which could be subdivided into two main concepts. The disparity between groups correlates with each group’s initial process of individual or group brainstorming with the former leading to more design concepts. However, overall each group generated a unique design concept, with each one differing from the others (Figure 2).

Prototyping Process
Two groups (Groups 1 & 2) began the prototyping session brainstorming together as a group and the third group began working independently and then eventually came together as a group to discuss their individual ideas. Interestingly each group had a slightly different process (Figure 3) that led to the generation of their final design concept. Despite this variation all three groups settled on their final design concept relatively early in the prototyping session and spent the rest of the time iterating and flushing out the design. All three groups had articulated their design concept prior to formally discussing the various requirements and data elements of the tool itself. Once the design had been articulated, the teams explored the data it could potentially present, how users could interact with it, and the goals it would accomplish. The teams spent much of their time discussing these various
requirements in no specific order, cycling between the topics and requirements looping back to their design, and repeating for the duration of the prototyping session. With each iterative conversation clarifying their concept as well as the problem the design addressed, and ultimately moving the group closer towards their final design.

A significant portion of the group design time was spent discussing various visual encoding strategies and how best to represent different data elements. The most common strategies discussed were color, size, position, and iconography. Similar to the primary design concept, once an encoding strategy was presented there was limited debate within the group if other strategies had better visual potential. Most of the discussion instead surrounded what data could be presented by the specific encoding method chosen that supported the overall design concept and framework.

All three groups explored familiar technological tools and applications, including both medical and non-medical references, which led to a significant influence on their final design. The largest influence, modern multi-touch devices, provided some of the most fundamental functional requirements for each team. In fact, we observed on multiple occasions the participants attempting to interact with their design as if it were a functional multi-touch tablet device, touching or swiping with their fingers. In addition to the interaction, design inspiration came from various smart phone applications including those focused on weather, social media, and other productivity tools. Finally, groups also turned to more familiar medical application tools, such as the current electronic medical record used at the study site and even flow cytometry. Groups had no trouble finding inspiration to help guide their overall design process.

In the second half of the prototyping session, once each group had a strong idea of their final design concept, the primary author (AP) went to each group asking them to discuss their design. Through open-ended as well specific questions, this interaction often uncovered an incomplete design or even idea. For example, in response to the question: “How does the tool highlight this important piece of information?” it wasn’t uncommon to hear the response: “Oh, I hadn’t thought about that yet” or “That’s a great question, we still need to figure that out!”

In addition, AP discussed with each group the potential advantages and disadvantages of the different encoding strategies utilized in the various designs. Teams were asked if they had (1) recognized the advantages and disadvantages of different encoding strategies and once recognized (2) if they had any concerns with their choices. Despite the groups employing encoding strategies with less precision than other techniques (e.g. area vs. length), we uncovered the motivation behind the encoding, and more importantly why they needed the data or information for their specific task.

**Individual Participation**

All individuals participated in the design session, contributing in all phases of the study. However, the senior physicians dominated the discussions in two of three groups as well as the group interaction. This influence was apparent in both the overall discussion as well as the design decisions individual groups made during the process. At no time did any of the researchers overhear or observe any behavior or actions exhibited by any of the physicians that could have been interpreted as creating a power hierarchy within a group. However, Group 1, which had the least variability in terms of participant seniority, had the highest level of equality regarding the degree of individual participation. As the discussion behind the design concept came to a conclusion and the groups transitioned to actually producing their final design drawings, participation within all the groups became more equally distributed.

Finally, two of the physicians had clinical responsibilities during the design session, which led to repeated interruptions, at times leaving for more than 5 minutes at a time. After returning, these two participants had potentially missed critical discussions and decisions, and had to spend time getting themselves caught up with the rest of the group.
The physician participants in our study not only identified valuable information to help design a future information tool used within the hospital setting, they also provided great insight into utilizing participatory design methods with a group of domain experts. Under the appropriate conditions, our physician designers demonstrated the effectiveness of PD as a design method, showing its potential to improve the technology utilized in healthcare organizations. In our discussion, we describe why PD works with physicians, some of the challenges we experienced and finally go on to identify a set of guiding principles for physician focused participatory design sessions.

Showing Knowledge-in-Action
Initially the brainstorming session might have been too vague for some participants, with a few needing additional instruction and direction. However, once started, the group maintained significant momentum carrying it forward allowing seamless transitions between activities. In addition, the warm up activity allowed the participants to become comfortable with each other as well as the design challenge itself. Given the number of unique and overlapping ideas as well as the speed at which they were generated, participants indirectly expressed their clear motivation to improve and advance the current state of health information technology.

During the brainstorming activity, participants generated many inter-related ideas quickly. Participants’ facility with the idea generation aspect of PD also demonstrates the high level of domain knowledge of our expert end users. The physician designers quickly and easily articulated their needs responding and building off each other’s ideas. Schon refers to this as “knowing-in-action,” which he defines as “the sorts of know-how we reveal in our intelligent action, publicly observable…like riding a bicycle and private operations like instant analysis of a balance sheet. In both cases, the knowing is in the action. We reveal it by our spontaneous, skillful execution of the performance, and we are characteristically unable to make it verbally explicit.” The physician designers’ ability to generate design concepts without first needing to develop functional requirements or explicitly stating a use case clearly supports the theory. Our designers used their implicit knowledge of their practice patterns and information needs to generate their initial design concepts. Through the PD process they had to explain and further expand on their original idea, unpacking their expertise and making it explicit. Not only did PD uncover and expose the underlying process of clinical prioritization, it also forced the participants to step back and reflect on their current practice ultimately defining their true needs. The activity required them to ask questions of their workflow and clinical processes that under regular circumstances do not require a significant amount of cognitive processing to execute as expert practitioners.

Throughout the session, participants engaged in progressive iteration, building off of their initial design concept as they worked. In some ways, participants’ approach resembles the practice of agile software development. Agile software development is a process “in which requirements and solutions evolve through collaboration…It promotes adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change.” Each of our groups followed a similar process where they generated an idea, iterated and brainstormed on that idea, ultimately incorporating it into their design, and then repeating this process for the duration of the design period. It is interesting to note that no groups followed a more traditional waterfall approach where specific requirements are identified and settled upon before moving onto the next phase of the design process. It is unclear if this relates back to their expert status, the participatory design process, or just their excitement to create and engage in the process, but the results clearly suggest agile methods as a successful tool for generating ideas with this population of experts.

Becoming PD-atriicians
The authors of this paper have collectively conducted and observed dozens of participatory design workshops over the years. One of us has considerable experience working with patients as designers, and another has conducted workshops with youth and teens. But the physicians in this study were a different breed. In some areas, they outperformed our expectations. For example, given relatively little guidance, participants in our study were able to generate sophisticated ideas during the individual brainstorming phase, and synthesize those into broad concepts during the whiteboard discussion. We were particularly struck by how little time and effort each group devoted to picking a design concept. In our experience with non-physician participants, participatory design groups will often spend significant time in this phase, trying out different design concepts and playing with the craft materials as a way to express competing ideas. With our physician designers, the opposite happened. Groups settled on an idea relatively quickly, and then used the materials as a way to realize their vision and express it for presentation. While selecting a single idea early certainly allowed participants to create more in-depth designs, we missed the
opportunity to hear additional concepts and ideas from participants as most groups did not discuss and debate different design ideas.

In their professional lives, our participants are frequently called on to synthesize data and quickly reach intuitive conclusions. They are often asked to improvise a solution based on limited data, but rarely called upon to design something or engage in blue-sky thinking. They are practitioners, and are used to working on deadline. Indeed, some of the effects we described above may simply arise from participants’ awareness of the tight schedule of the design session, and their desire to ensure they had something to present. Finally, the physicians were articulate presenters and offered collegial and supportive critique.

Nevertheless, our study shows that participatory design with physicians can work well, provided the workshops themselves are designed with physicians in mind. Brainstorming and ideation can be accomplished successfully with relatively little guidance, provided the intended design is one that meets physicians’ current needs and experiences. Designs may not shift dramatically from early concepts, but a participatory design exercise can still capture value through surfacing physicians’ priorities. The participants were clearly engaged by the exercise and very motivated to solve real information challenges they experience while providing clinical care to their patients.

Recognizing Challenges and Constraints
Despite the overall success of the design session, we also experienced several challenges. We had no trouble finding interested participants; instead, we had difficulty finding a time to conduct our two-hour design session that fit into the busy schedule of our physician designers. Initial efforts to conduct a session during regular office hours failed to find a time that worked for multiple participants. Therefore, we elected to hold an evening dinner session, scheduled one month in advance, which made it much easier for our physician designers to participate. In addition, because we conducted the session after hours, participants seemed to have fewer distractions (checking their email) and less time constraints. In the end, the timing of our session allowed the participants to free themselves of their usual routines and fully immerse themselves in the design process.

Organizing the design sessions, within the two-hour time constraint, had its own challenges. We needed to ensure adequate time for brainstorming, designing, sharing, and ultimately feedback. Even though the design exercise took up over half of the entire session, having more time might have benefited both the physician designers as well as the research team. We asked that the physician designers produce one final design concept for each group, but did not put any methodological constraints on how to reach that final design. As we described, the teams settled on their design relatively early and refined the concept for the majority of the session. Rapid serial iteration has been shown to increase the diversity of design concepts leading to more successful products. In addition, producing multiple alternative ideas in parallel encourages designers to uncover previously unnoticed constraints and identify new opportunities. Unfortunately, given the time constraint of the session, we did not require or even encourage the physician designers to develop multiple ideas in parallel. Even Group Three, which spent time individually brainstorming ideas, only focused their efforts on a single design. In addition, people are more likely to provide more honest and critical feedback when they present and are presented with multiple designs versus single ideas.

When our physician designers shared their final concepts, the other teams for the most part provided positive feedback. We clearly saw the influence of social dynamics and professional hierarchy play out during other parts of the session. Senior physicians typically spoke first, for longer periods of time, and often led the conversation in each of the groups. When assembling the groups, we attempted to minimize this phenomenon, though we had limited success. In addition, the groups that exhibited more obvious participation discrepancies tended to delegate the less cognitive tasks to the more junior physicians. With idea generation one of the primary goals of participatory design, it is important to understand and recognize the role social dynamics might play in influencing productivity within a group of physician designers.

The final issue we encountered concerned the actual designs themselves. Because the physician designers had no or limited design experience, some of their designs incorporated visual encoding strategies that went against well-defined best design practices. Designs attempting to communicate large volumes of numerical information should leverage the innate ability of our visual system to recognize certain information without any higher-level cognitive input. Pre-attentive processing allows for rapid interpretation of visual information with the proper

Figure 4: Leveraging pre-attentive processing. Utilizing the innate ability of our visual system allows individuals to identify differences in length more accurately than those of area. In both examples the shape on the right is 5% smaller than the shape on the left.
framing and organization of the information. Size has the potential to communicate information rapidly, leveraging pre-attentive processing, though only when done correctly. For example, it is much easier to recognize size discrepancies along a single dimension, such as length versus a difference found in two or more dimensions, such as area (Figure 4). Some of the final designs produced during the prototyping session utilized differences in size along two dimensions (i.e. area) to communicate differences in scale. As designers we immediately recognized the potential limitation of this encoding strategy, though when pressed the physician designers seemed less concerned. However, while the physical designs created during the PD sessions have their place in the final design, more important are the ideas, values, and needs expressed by the non-expert designers. The output of participatory design is not a final design specification, but rather information for an expert designer to use when designing the final product, taking into consideration the physical design as well as the needs and requirements identified during the design session by the domain experts.

Engaging PD-atricians
When we step back and evaluate our own process, the PD session, as well as the actual content generated, we identified an underlying theme that led to the overall success of our work. Having a facilitator with a detailed understanding of the organization, domain expertise in the clinical context as well as human centered design methods allowed us to tailor and adapt the sessions to ensure they achieved the desired outcomes. With this knowledge, we identified a motivating challenge and recruited participants with the potential to make a meaningful contribution to the PD process. In addition, since the lead author was a peer of the session participants, it allowed them to trust this new, unfamiliar process and engage in a way that would not have been possible without the professional connection. Therefore, organizations investing in HIT and a desire to employ PD methods need to ensure they have informatics professionals with these diverse skill sets and established interpersonal relationships.

Guiding Principles for Future PD-atricians
Given the high costs associated with health information technology system failures, organizations need to take the time required to ensure the technology leads to the desired benefit and not introduce any harm. Although some researchers have described the economic benefits of iterative design and others have clearly demonstrated its rigor, many organizations simply undervalue its potential. For less than $500 (costs associated with this PD session), we now have a solid foundation to begin the design process. In addition, by engaging physicians in this process, not only do we gain their valuable insight, but also their interest in seeing the successful deployment of the tool as they are now partial owners of its ultimate success. In order to support others as they pursue participatory design methods to help advance health information technology, we offer a series of guiding principles developed from our presented experiences (Table 3).

In addition to the principles and ideas previously discussed, it is important to ensure that appropriate subjects are selected to participate in the PD session. Ideally, the organizers want to find those who are interested and acknowledge the existence of the challenge or problem. Participants need to have good communication skills, be open to feedback, and be willing to engage in the design process, as well as domain expertise. While it is important to have a diverse group of participants, session organizers need to consider the potential for these differences to negatively influence the group’s performance. Therefore, playing close attention to how the groups are structured should help to minimize any negative group dynamics.

Besides needing a diverse group of participants, it is just as important to have a diverse team facilitating the PD session. In the medical setting, the team needs to have knowledge and expertise not only in PD and design principles, but also the domain itself. Having these backgrounds allows the facilitators to ask directed and focused questions with the intent of pushing the participants to expand their thinking and ultimately their designs. The diversity of our backgrounds allowed us to form questions based on both the domain content as well as design aesthetics. Our team also had the benefit of our primary author being a physician and peer of the session participants, with a proven track record of improving the clinical information tools at the study location. The personal connection with each of the participants improved our recruiting capabilities as well as lead to increased engagement.

Table 3: PD-atrician Guiding Principles

<table>
<thead>
<tr>
<th>Guiding Principles</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Include a diverse set of domain and process experts for both organizers and participants</td>
<td></td>
</tr>
<tr>
<td>2. Identify a motivating challenge</td>
<td></td>
</tr>
<tr>
<td>3. Support iteration and agile methodologies</td>
<td></td>
</tr>
<tr>
<td>4. Allow for reflection and discovery</td>
<td></td>
</tr>
<tr>
<td>5. Require multiple designs</td>
<td></td>
</tr>
<tr>
<td>6. Identify and plan around time constraints</td>
<td></td>
</tr>
<tr>
<td>7. Select participants and groups thoughtfully</td>
<td></td>
</tr>
<tr>
<td>8. Create a safe environment, to foster trust with the organizers and the PD process</td>
<td></td>
</tr>
<tr>
<td>9. Recognize that the ideas, values, and motivation behind a specific design is just as important as the design itself</td>
<td></td>
</tr>
</tbody>
</table>
Limitations

Despite the diversity of medical providers within our sample, all of the participants were pediatricians at a single organization. In addition, Seattle Children’s Hospital supports the active engagement of its medical providers and encourages their participation in continuous process improvement activities. Therefore, they are very familiar with identifying system challenges and working collaboratively and iteratively towards consensus driven solutions. These factors might limit the generalizability of our study since physicians at other organizations may be less familiar with these organizational processes. However, given PD methods support non-experienced designers in the creation process, it should minimize these concerns.

In addition, we need to recognize the fact that only 11 individual physicians participated in a single PD session. While this also raises questions about generalizability, we feel our findings still provide insight into a process with genuine potential. Strauss and Corbin, describe the value of recognizing both the messages or theories derived from qualitative work, as well as the situation and conditions from which the theory was derived. Therefore, with the provided descriptions of our methods and institutional culture, others can use our work as a starting point and adapt it to fit their own organization.

Conclusion

Participatory design, a method well known to the Human-Computer Interaction community, has the potential to significantly enhance and improve the development of health information technology. Shifting design activities that involve end users to earlier in the development process should lead to applications that better meet the needs of stakeholders. Though it has seen limited adoption in the healthcare community, we have demonstrated its potential to make significant contributions. Utilizing well-described methods, we provide a series of guiding principles to engage physicians in designing new clinical information tools through participatory design. Organizations need to invest in human centered design methods such as PD by training their existing informatics professionals. Utilizing practicing clinicians with experience in Human-Computer Interaction, informatics, and deep domain expertise, project teams can build trust with typically skeptical end users leading to active participant engagement, translating into more successful projects.

References

13. Card SK, Mackinlay JD, Shneiderman B. Readings in Information Visualization: Using Vision to Think.


Resource Classification for Medical Questions

Kirk Roberts, PhD\textsuperscript{1}, Laritza Rodriguez, MD, PhD\textsuperscript{2}, Sonya E. Shooshan, MLS\textsuperscript{2},
Dina Demner-Fushman, MD, PhD\textsuperscript{2}

\textsuperscript{1}School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX
\textsuperscript{2}Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD

Abstract

We present an approach for manually and automatically classifying the resource type of medical questions. Three types of resources are considered: patient-specific, general knowledge, and research. Using this approach, an automatic question answering system could select the best type of resource from which to consider answers. We first describe our methodology for manually annotating resource type on four different question corpora totaling over 5,000 questions. We then describe our approach for automatically identifying the appropriate type of resource. A supervised machine learning approach is used with lexical, syntactic, semantic, and topic-based feature types. This approach is able to achieve accuracies in the range of 80.9\% to 92.8\% across four datasets. Finally, we discuss the difficulties encountered in both manual and automatic classification of this challenging task.

Introduction

The answers to medical questions can be found in a wide variety of resources. These resources include general medical knowledge (such as that found in textbooks or encyclopedias), highly specialized research knowledge (such as the biomedical literature), and even source specific to the given patient (such as the patient’s electronic health record (EHR) information). For automatic question answering (QA) systems that intend to answer a wide variety of medical questions, having the ability to properly target a specific resource (such as a corpus, sub-corpus, or database) would be very useful. Consider the following three questions from the datasets used in this study:

(1) \textit{What kind of allergy does he have?}

(2) \textit{Is it necessary to wait until age 7 before doing an allergy evaluation?}

(3) \textit{Is there anything in the literature about allergy testing or desensitization to acyclovir?}

While all these questions are topically concerned with allergies, they require different resources for answers. The first question is patient-specific, and the answer must come from the patient’s allergy list or notes in the health record. If the information is missing there, the patient himself must be questioned, as any attempt to look in a non-patient-specific source will yield useless results. The second question requires general medical knowledge, either in a textbook, practice guidelines, or manuals. The final question specifically asks for scientific literature, and would not be appropriately answered by a background knowledge data source. All three types of questions are commonly asked by clinicians at the point of care (indeed, all three of the above questions were asked by physicians and collected by Ely\textsuperscript{1} and Patrick and Li\textsuperscript{2}). A QA system designed to aid clinicians at the point of care, therefore, should be able to determine which resource contains the most appropriate answer to the question.

In this work we develop both manually annotated datasets and automatic methods for classifying questions by their intended resource. We focus on the three types of resources discussed above: patient-specific, general or background medical knowledge, and scientific research. To the best of our knowledge, this is the first work to address this particular facet of QA despite it being a critical task in developing general-purpose QA systems for clinicians. We have manually annotated over 5,000 questions from four existing question collections. A supervised machine learning method is then employed to classify questions using lexical, syntactic, semantic, and topic-based features. Finally, we discuss the difficulties in both manual and automatic resource classification for questions.
Background

The need for medical QA systems to handle a wide variety of questions has resulted in questions being classified along many dimensions: answer type, topic, relation, user, answerability, and now resource. These will be discussed in more detail below, but the type that impacts the most on resource choice (besides what we discuss in this paper) is the user. Specifically, the assessment of the user’s medical expertise. Users are often organized into one of two groups: consumers and professionals\cite{3,4} (though it has been shown that this distinction is oversimplified\cite{5}). Importantly, resource choice is based on the user profile\cite{6}; while a physician might be directed to a resource such as UpToDate, a consumer might be best suited by MedlinePlus. This is largely orthogonal to the study of resource type presented here, however, we acknowledge that certain general knowledge sources are better suited to consumers, while others are better suited to professionals. The same might be said about patient-specific and research resources, where patient portals and consumer-focused research news sites such as HealthDay\footnote{http://consumer.healthday.com/} might be considered more appropriate for consumers. Thus, resource selection requires both understanding the user and type of knowledge source. For the purpose of this study, we focus entirely on the resource type.

As noted above, there are many other ways of classifying questions. It is important to understand the breadth of medical question classification schemes, as many have an impact on resource classification. The most common classification in the open domain is answer type, sometimes referred to as the expected answer type or even simply question type. The answer type is the semantic type of the answer, and is useful in candidate answer extraction. Answer types are more studied in non-medical questions\cite{6,7,8}, where factoid-style questions with entity answers are common. Answer type methods for medical questions have been proposed by Cruchet et al.\cite{9}, who studied a small number of English and French questions, and by McRoy et al.\cite{10}, who focus on cancer-related questions. In analyzing answer types, both Cruchet et al.\cite{9} and Roberts and Demner-Fushman\cite{4} have found that boolean (i.e., true/false, yes/no) questions are the most common surface type. However, as observed by those researchers, often a higher-level type is more appropriate. For example, the question “Can Group B streptococcus cause urinary tract infections in adults?” is a boolean question on the surface, but is really asking for cause/etiology. This higher level answer type has been called many names: medical type, topic, and commonly just question type. This is the most common type of question classification studied by researchers, and has a critical impact on resource as some types are more likely to be discussed in some resources than others (e.g., definition questions are appropriate for general knowledge resources, while comparative effectiveness of treatments is more likely to be discussed in research studies). Among the proposed methods for this general question type are Cruchet et al.’s bilingual system, Roberts et al.’s\cite{11} consumer QA system, and several systems utilizing Ely’s\cite{12} general topics\cite{13,14,15}, which includes types such as diagnosis, history, prognosis, and treatment/prevention. Many medical QA approaches have focused on identifying question templates, common question types with slots for specific items with the question. Cimino et al.\cite{16} maps questions to templates such as “PHARMACOLOGIC\_SUBSTANCE treats DISEASE\_OR\_SYNDROME” using a rule-based approach. Patrick and Li\cite{2} developed EHR question templates such as “Did the patient have X, Y?” and utilized a multi-layered supervised machine learning approach to classify templates and slot components. A very common template approach for research questions is PICO (problem/population, intervention, comparison, and outcome). The PICO approach has been utilized by Demner-Fushman and Lin\cite{17} and Schardt et al.\cite{18}, amongst others. Beyond these types of question classification, other approaches have studied (a) relation extraction\cite{19} using SemRep\cite{20}, (b) disease classification\cite{21}, (c) anatomy classification\cite{13}, and (d) question answerability\cite{22,23}.

As mentioned, all these types of question classification have a relation to the best resource (patient-specific, general, research) in which to find the question’s answer. The previous work that is the most overtly overlapping with our own are the answer type approach by McRoy et al. and the answerability approaches by Yu and colleagues. McRoy et al.’s\cite{10} answer types form a taxonomy that contains three top-level nodes: factual, patient-specific, and non-clinician. Patient-specific questions correspond to those considered here, while factual questions may include both general knowledge and research. Non-clinician questions deal with non-medical issues such as insurance and legal advice. We found such questions in our dataset as well, but exclude them from our study as we are focused on clinical questions. Similarly, the answerability classification performed by Yu and Sable\cite{22} and Yu et al.\cite{23} isolates patient-specific questions as “unanswerable” based on Ely’s taxonomy\cite{12}. In all these cases, no distinction is made between research questions
and general/background medical knowledge. Admittedly, this distinction is often unclear, so in the next section we attempt to develop an initial specification for when questions belong to different resource categories.

Data and Annotations

We utilize four different question sets:

1. **ELY** ($n = 1,500$): These questions come from the Clinical Questions collection$^2$ which was collected by Ely et al.$^{[1,24,25]}$ and D’Alessandro et al.$^{[26]}$ and is maintained by the National Library of Medicine (NLM). The questions were collected from physicians, either during direct observation or during a phone interview. They are largely designed to represent the stream-of-conscious questions that physicians have on a day-to-day basis, the vast majority of which go unanswered. There are 4,654 questions in the collection, from which we randomly sampled 1,500.

2. **MEDPix** ($n = 1,666$): These questions came from MedPix$^3$, which is an image database with over 53,000 images from over 13,000 patients. Originally built by James Smirniotopoulos at the Uniformed Services University, it is now maintained by NLM. MedPix contains hundreds of cases that are associated with quiz-style multiple choice questions, allowing radiologists to receive CME credits. The question set contains patient-specific questions forcing users to review the patient’s image and note, as well as medical knowledge questions inspired by the patient but not requiring any patient-specific answer. These knowledge questions may involve general radiology knowledge, or require being familiar with recent research.

3. **GARD** ($n = 1,476$): These questions were submitted by consumers to the Genetic and Rare Diseases Information Center (GARD)$^4$. The questions are then curated, answered by NIH experts, and maintained on the GARD website in an FAQ format. Despite being consumer questions, the nature of the subject matter makes them more similar to clinician questions than other types of consumer questions$^5$. The questions were originally in a paragraph style, but manually decomposed into individual questions$^6$. When isolated from the original request, the questions appear as either patient-specific (as they are about details provided in the full request) or for medical knowledge. Since they are topically concerned with genetic and rare diseases, the users mostly understand the scientific literature might be the only source of answers for the particular disease.

4. **Li** ($n = 486$): These questions were collected by Patrick and Li$^7$ and are included in Li’s dissertation$^8$. They are organized by the taxonomy discussed in Patrick and Li. The focus of the questions are patients in the intensive care unit, and thus almost all are patient-specific. We use only the sub-set of questions, defined by the taxonomy, of clinical relevance (ignoring, e.g., questions about the hospital organization). As such, they need not be manually annotated in the manner of the above corpora. Instead, the existing taxonomy classification is utilized, where only a small number (7) of general information questions are included.

A selection of questions from these corpora can be seen in Table 2.

To annotate each question, we used the following guidelines to distinguish between patient-specific, general, research, and other questions.

a. **PatientSpecific**: In these questions, the answer is either contained directly in the patient’s chart or information needs to be retrieved from the chart to answer the question.

- What is her latest A1c value?
- Is this rash shingles or staphylococal impetigo?

---

$^2$http://clinques.nlm.nih.gov/
$^3$http://medpix.nlm.nih.gov/
$^4$https://rarediseases.info.nih.gov/gard

1042
Table 1: Annotation data for the four corpora. The LI corpus was annotated using data from Li[28].

b. GENERAL: In these questions, the answer should be contained in a general medical knowledge source, exemplified by a medical textbook. This could include textbooks for specialties and sub-specialties.

- What causes scleromyxedema?
- How do you do a paracentesis?
- What are the immunizations for an 18-year-old?
- What urological symptoms are associated with Beckwith Weidemann Syndrome?

c. RESEARCH: In these questions, the answer should be best found in a research-type source. This includes both original scientific articles (bench and clinical) and articles summarizing the state-of-the-art in some area, including reviews and practice guidelines. While guidelines could arguably be considered general knowledge, we include them as research because they are typically updated more frequently than textbooks and they are closely associated with review articles that are published in the scientific literature.

- What research is being done at present regarding CDPX1?
- What percent of patients with Congenital Adrenal Hyperplasia (CAH) will have upper genitourinary tract abnormalities (e.g., vesicoureteral reflux and hydronephrosis)?
- How are the newer medications for attention-deficit hyperactivity disorder being used?
- Is glimepiride better than other sulfonylureas for diabetes?

d. OTHER: These are non-clinical questions. Sometimes they are financial in nature, other times they are marked as OTHER because it is too difficult to understand the question and they should not be considered part of the dataset.

- Mother paid $101 for 50-dose desmopressin spray. Can that be the correct cost?
- Can you help me?
- Which of the following is True?

All of the questions except the LI set were double-annotated by an MD/PhD (LR) and a medical librarian (SS). Table 1 contains the reconciled statistics for all four corpora: numbers (and percents) of PATIENTSPECIFIC, GENERAL, RESEARCH, and OTHER questions. As can be seen, the three manually annotated corpora are heavily imbalanced toward general questions (one of the main reasons we chose to also include the almost entirely patient-specific LI corpus). Table 1 additionally contains the κ agreement for the three manually annotated corpora, ranging from 0.62 to 0.78. These inter-annotator agreement numbers are decent, but by no means great. This should serve to emphasize that resource choice is still difficult for humans. Half of the disagreements were GENERAL versus RESEARCH (ELY: 50%, MEDPIX: 51%, GARD: 41%). Many of the disagreements involved confusion of what to consider OTHER (ELY: 34%, MEDPIX: 26%, GARD: 50%), so we had to carefully re-evaluate what to consider part of the scope of this work. Most of these OTHER disagreements were due to peculiarities in the datasets themselves and so should not be a concern for a clinical QA system. For instance, the GARD questions were automatically extracted from longer questions, resulting in some elliptical questions such as “How?” and “Can you help me?”. In MEDPIX, some of the questions made little sense without knowing the multiple-choice answers. These are not the type of question to be asked without context to a clinical QA system, so these inter-annotator disagreements can safely be ignored.
What kind of allergy does he have?

(a)

What

allegy

does

he

have

(b)

case

of

kind

have

does

he

(c) (d) (e)

Methods

Here we present an initial approach for automatically classifying the resource type for medical questions. Our goal is to characterize the linguistic criteria necessary to distinguish between questions requiring different resource types. As such, we forgo exhaustive experimentation and instead experiment with a variety of features at different linguistic levels: lexical, syntactic, semantic, and topic.

1. **Lexical**: We consider two bag-of-words features and length-based lexical features. The bag-of-words features include a simple unigram feature and a subset of unigrams that do not belong to a medical concept. To recognize concepts, we utilize the latest version of MetaMap$^{[29]}$ and restricted concepts to a specific set of UMLS$^{[30]}$ semantic types: `diap`, `cgab`, `acab`, `inpo`, `patf`, `dsyn`, `anab`, `neop`, `mobd`, `sosy`, `drdd`, `clnd`, `antb`, `phsu`, `nsba`, `strd`, `vita`. Thus, this feature could be considered lexico-semantic as it incorporates a semantic resource. The purpose of the non-concept unigram feature is to ignore medical terms that may be used in any resource (such as diseases) and instead focus on words that help delineate resource type. Additionally, we experiment with word- and character-length features. Upon observation, RESEARCH questions tend to be longer (in words) and contain words that are longer (in characters) than GENERAL questions.

2. **Syntactic**: We consider a syntax feature based on the syntactic dependency parse. Syntax can help disambiguate different uses of the same word. For example, the word `patient` in “What is the diagnosis for the patient” (PATIENTSPECIFIC) and “My patient has Hepatitis A and I’m wondering if there are any recent treatments?” (RESEARCH). We use the Stanford CoreNLP dependency parser$^{[31]}$. As features, we consider subgraphs of the full dependency parse. Figure 1 shows a sample of dependency subgraphs for the question “What kind of allergy does he have?” Since these are even sparser than n-grams, we pre-calculate the most frequent subgraphs in each question corpus using the ParSeMiS implementation$^5$ of the Gaston frequent subgraph mining algorithm$^{[32]}$ similar to the method employed in Luo et al.$^{[33]}$. This still results in a large number of subgraphs, so we further filter based on the statistical association of a subgraph and an output class using Fisher’s exact test, taking all subgraphs above a threshold. All Fisher scores are calculated on the training set.

3. **Semantic**: We consider a concept feature based on MetaMap as well as distributional semantic features that represent each question in a semantic vector space. The concept feature indicates the semantic types present in the question, as different semantic types are more prevalent with different resource types. The distributional semantic features utilize paragraph2vec$^{[34]}$, a word sequence extension to the popular word2vec word

5https://github.com/tintad FOLLOWING
### Table 2: Most similar questions using question-level embeddings and LDA-based topics.

<table>
<thead>
<tr>
<th>Question 1</th>
<th>Does he have spleen lacerations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embeddings</td>
<td>PATIENTSPECIFIC</td>
</tr>
<tr>
<td>1. Does she have ongoing haemolysis?</td>
<td>PATIENTSPECIFIC</td>
</tr>
<tr>
<td>2. Is he ventilated?</td>
<td>PATIENTSPECIFIC</td>
</tr>
<tr>
<td>3. Did she have sepsis overnight?</td>
<td>PATIENTSPECIFIC</td>
</tr>
<tr>
<td>Topics</td>
<td></td>
</tr>
<tr>
<td>1. Does it matter whether you draw the prostate specific antigen before or after the rectal exam?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>2. What causes polyembryoma?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>3. How are they treated?</td>
<td>GENERAL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question 2</th>
<th>What causes scleromyxedema?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embeddings</td>
<td>GENERAL</td>
</tr>
<tr>
<td>1. What causes pilomatrixomas?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>2. What causes nonpalpable purpura?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>3. What causes excessive sweating?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>Topics</td>
<td></td>
</tr>
<tr>
<td>1. What is his neurological status?</td>
<td>PATIENTSPECIFIC</td>
</tr>
<tr>
<td>2. I am interested in finding out information related to dihydropyrimidine dehydrogenase (DPD) in children.</td>
<td>GENERAL</td>
</tr>
<tr>
<td>3. What is the differential diagnosis of neonatal seizures?</td>
<td>GENERAL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question 3</th>
<th>Is there anything in the literature about allergy testing or desensitization to acyclovir?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embeddings</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>1. Is there anything in the literature about milk and molasses enemas?</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>2. Which of these features of a &quot;ring lesion&quot; is least common in an abscess?</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>3. Is there anything in the literature on elbow fat pad signs?</td>
<td>RESEARCH</td>
</tr>
<tr>
<td>Topics</td>
<td></td>
</tr>
<tr>
<td>1. Is there a genetic test which can help to rule-out neurofibromatosis?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>2. Is a sample for genetic testing are required?</td>
<td>GENERAL</td>
</tr>
<tr>
<td>3. Any information you have about any current therapies could be helpful.</td>
<td>GENERAL</td>
</tr>
</tbody>
</table>

Table 3: Automatic resource type classification results across four corpora.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>ELY</th>
<th>MEDPIX</th>
<th>GARD</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Frequent Class</td>
<td>76.27%</td>
<td>77.06%</td>
<td>86.14%</td>
<td>71.88%</td>
</tr>
<tr>
<td>Unigram Bag-of-words</td>
<td>78.97%</td>
<td>91.34%</td>
<td>90.69%</td>
<td>78.83%</td>
</tr>
<tr>
<td>Best feature set</td>
<td>80.85%</td>
<td>92.76%</td>
<td>91.13%</td>
<td>79.28%</td>
</tr>
</tbody>
</table>

The gensim implementation of paragraph2vec (known there as doc2vec) is used with default options. The vectors are pre-built using all questions as well as the 450,000-sentence health reference corpus used in Kilicoglu et al.[37]. We experiment with two versions of this feature: the first uses the embedding vector directly, while the second counts the output classes of the k-nearest neighbors (e.g., 5-NN) using cosine distance between the embedding vectors on the questions from the training set. Table 2 shows question similarity examples using embeddings based on PARAGRAPH2VEC.

4. **Topics**: We consider topic model features that represent questions as distributions over topics. We utilize the gensim version of Latent Dirichlet Allocation[38]. The unlabeled data and features are otherwise identical to the embedding features described above. Table 2 also shows question similarity examples using topics based on LDA.

Using the above features, we utilize a single 3-class support vector machine (SVM) with a linear kernel[39]. Four datasets are used for evaluation: (1) ELY, (2) MEDPIX, (3) GARD, and (4) ALL = {ELY, MEDPIX, GARD, LI}. The LI dataset alone is too imbalanced for consideration. The feature set for each dataset is chosen from the features described above using a greedy search algorithm[40] on a development split of the data. The evaluation results reported in the next section are computed using a 5-fold cross validation of each corpus.

**Results**

The selected features for each dataset are as follows:
As can be seen, different features were automatically selected for each dataset. A unigram feature (either the basic unigram or the non-concept unigram feature) was chosen as a base feature, while lexical, syntactic, semantic, and topic features were chosen to augment the bag-of-words feature. No dataset utilized the distributional semantic features, despite these being a good method for assessing similarity. It is thus likely that the distinguishing factor between resource type questions is quite localized (a single word or short phrase), while the similarity based on paragraph2vec is based on the entire question. In our experiments, the paragraph2vec-based vectors tended to better represent overall question similarity than the LDA-based vectors; however, the latter were chosen in three datasets (ELY, GARD, and ALL). Since LDA uses a bag-of-words assumption instead of the language model type assumption used by paragraph2vec, it tends to group words that are used in the same context but have very different semantic functions (this is why it is referred to as a topic model). This topical grouping effect might be more robust for this task than the full-question semantic similarity used by paragraph2vec.

Table 3 contains the results for each dataset. The baseline unigram feature alone greatly outperforms the weak baseline of choosing the most frequent resource type (ELY: +2.7%, MEDPix: +14.28%, GARD: +4.55%, ALL: +6.95%). The selected features, however, only slightly outperform this stronger baseline feature (ELY: +1.88%, MEDPix: +1.42%, GARD: +0.44%, ALL: +0.45%). This suggests that not only are simple lexical features the most important for classifying resource type, but the higher-level linguistic features are not easily representable with the features studied here (despite the fact that these are state-of-the-art features in many text classification tasks). In the Discussion below, we perform some error analysis to uncover the links between human disagreements and errors in the automatic system.

**Discussion**

Providing relevant answers ultimately depends on finding the most appropriate source of information for a given question. As discussed in the Background, this involves not only the semantics of the question itself, but also the user’s background knowledge. In this work, however, we seek to isolate the resource choice from other considerations, grouping resources in a cross-sectional way (e.g., all general medical knowledge resources, regardless of whether the user is a high school student or highly-trained cardiologist). This reduces the need to jointly understand all the possible resource choice factors at once, which is especially useful if a system is designed to be used by a specific sub-group (e.g., clinicians).

At the highest level of granularity, the resources can be split into **PATIENTSPECIFIC** (e.g., EHR or personal health record), **GENERAL** (e.g., textbooks and other background knowledge literature), and **RESEARCH** (e.g., journal articles, systematic reviews). Our results show that separating the literature (GENERAL and RESEARCH) into distinct buckets might not be feasible. Instead, one could consider a continuum ranging from the most basic background knowledge
(e.g., *Harrison’s Principles of Internal Medicine*) to the most late-breaking research (e.g., the latest updates on the Zika virus in PubMed). While some questions are unambiguously background knowledge (e.g., “Is Williams syndrome inherited?”), and others unambiguously require research answers (e.g., Question (3)), many questions fall into a zone of ambiguity. In terms of manual annotation, the GENERAL vs. RESEARCH ambiguity resulted in many disagreements. If one only considers questions both annotators agree are either GENERAL or RESEARCH (i.e., ignore questions either annotator marked as PATIENTSPECIFIC or OTHER), then inter-annotator agreement would be lower on two of the three datasets (ELY: 0.47, MEDPix: 0.28, GARD: 0.67). In terms of automatic annotation, this was a large source of classifier error, such as the following mis-classifications:

(4) Should a gestational diabetic have any antenatal fetal well-being testing done?

(5) What are the best exercises for rotator cuff?

(6) What is the success rate of various methods of endometrial biopsy? (that is, getting in)?

(7) Is Keflex the drug of choice for a patient with positive streptococcal screen (allergic to penicillin, sulfa)?

Here, Questions (4) and (5) are GENERAL questions mis-classified as RESEARCH, while (6) and (7) are RESEARCH questions mis-classified as GENERAL. The choice of resource type is largely based on the granularity of information needed. The annotators felt that the first two questions concerned large populations and did not ask for detailed specifics, while the second two questions asks for details that often don’t make it into textbook-style knowledge resources (i.e., the exact success rate for (6) and the review of efficacy of various streptococcal drugs for (7)).

On the other hand, PATIENTSPECIFIC questions are distinct and less ambiguous since they do not belong on such a continuum. However, there were difficulties when a question requires both patient-specific information and possibly external knowledge as well. For example:

(8) What should I do about this bullous lesion found on sinus films?

Here, more knowledge from the patient chart is needed regarding the lesion, after which other knowledge sources might be required as well. We chose to consider such questions as PATIENTSPECIFIC, but another option would involve a multi-label approach that might also label the question as GENERAL. In fact, this question was indeed mis-classified as GENERAL. Optimistically, however, one might claim that with an automatic QA system, users would not ask questions that required multiple information sources, so these questions should be less of a concern.

Perhaps, instead of using a 3-class resource type representation, the most ideal solution would be a large corpus of question-answer pairs. That is, instead of pre-deciding where a given answer is most likely to be found for a question, search a range of knowledge resources and identify the ones that actually contain an answer. The corpus created here was annotated in isolation from any possible answers (i.e., annotators did not look for an answer it multiple sources to see where one happened to be available). Rather, the annotators used their best judgement about where certain types of medical information would likely be found. With a large corpus of question-answer pairs, however, there is no need for such assumptions. The trade-off is obviously the additional cost of annotating answers over simply a 3-class determination in isolation. Our experience has indicated that annotating answers is incredibly time consuming[41], requiring approximately two orders of magnitude more time to find an answer than to simply classify a question in isolation. Further, rather than a 3-class classification problem, this would result in a far more granular representation which would either require far more annotations to achieve the same result, or an approach entirely different from the one presented here. Alternatively, a compromise solution could involve a more fine-grained hierarchical ontology of medical knowledge sources, encompassing many of the different types of publications (e.g., original research, systematic reviews, web articles). This could extend, for example, the MeSH article types to include other types of medical knowledge sources. Ambiguities could then be represented by selecting multiple nodes in the ontology.

**Conclusion**

We have presented our approach for classifying resource types of medical questions to aid QA systems in selecting the most appropriate data source. Three resource types were considered: patient-specific, general knowledge, and research
knowledge. To develop the approach, four different corpora totaling over 5,000 questions were annotated with the three resource types, with moderate inter-annotator agreement. An initial automatic method was proposed utilizing state-of-the-art features, including dependency subgraph, distributional semantic, and topic modeling features. Nonetheless, basic lexical features provided the vast majority of the gain, suggesting that many existing state-of-the-art features cannot easily distinguish between these questions. Further advances, therefore, are needed not only in the automatic classification, but also in the representational methods used to identify the most appropriate medical resource.

**Data Availability**  The annotated question corpora are available by request (kirk.roberts@uth.tmc.edu).

**Acknowledgements**  This work was supported by the National Library of Medicine (NLM) grant 1K99LM012104 (KR), as well as the intramural research program at NLM (LR, SS, DDF).

**References**

41. A Deardorff, K Masterton, K Roberts, H Kilicoglu, and D Demner-Fushman. A protocol-driven approach to automatically finding authoritative answers to consumer health questions in online resources. *In Submission*. 

1049
Acceptability of Fitbit for physical activity tracking within clinical care among men with prostate cancer

Dori Rosenberg, PHD, MPH1, Elyse A Kadokura, BS2, Erin D Bouldin, MPH, PHD2,3, Christina E Miyawaki, MSW, PHD4, Celestia S. Higano, MD5, Andrea L. Hartzler, PhD1

1Group Health Research Institute, Seattle, WA; 2University of Washington, Seattle, WA; 3VA Puget Sound Health Services Research & Development, Seattle, WA; 4University of Houston, Houston, TX; 5Fred Hutchinson Cancer Research Center

Abstract
Prior research has not examined the acceptability of commercially available fitness tracking devices in men with prostate cancer, many of whom are at risk for conditions that physical activity could alleviate. We conducted an exploratory 3-week field study to examine acceptability of the Fitbit Zip and attitudes towards integrating fitness tracking into clinical care among men with prostate cancer. Twenty-six men used the Fitbit Zip for a one-week baseline phase followed by a 2-week optional use phase and then completed in-depth interviews. Interview data was analyzed using inductive thematic analysis. Participants found the device comfortable and easy to wear. Barriers to use included health and technology difficulties. Participants expressed value in sharing Fitbit data with their health care team. Findings support the use of easy to use and simple fitness trackers among men with prostate cancer and there could be opportunities to integrate fitness tracker data into clinical care.

Introduction
Men with prostate cancer, particularly those taking androgen deprivation therapy (ADT), are at high risk for side effects of treatments such as loss of muscle mass and strength, increased fat mass and cholesterol, glucose intolerance, and decreased bone mineral density1,2. The negative impacts of these side effects can be mitigated by regular light or moderate-to-vigorous physical activity. There is evidence suggesting that men with prostate cancer who are provided with supervised and/or home-based aerobic and resistance exercise programs have improvements in quality of life, fitness, body composition, lower body strength, and glucose intolerance5,6. Exercise has also been associated with reduced risk of recurrence and mortality in men with prostate cancer5,6. In spite of these benefits, men with prostate cancer have objectively lower physical activity than the general population of men over age 607.

Patients with chronic disease have reported benefits from tracking health indicators such as weight, physical activity, and diet.8 Commercially available wearable activity trackers such as the Fitbit® are promising tools that could be used as a mechanism to encourage daily physical activity among populations that are highly inactive and could be employed to improve healthy aging in the general population9,10. They also have the potential to be used by health care providers as a way to monitor patient prognosis and recovery through tracking activity levels and mobility11. In a qualitative study, health care providers who were asked questions about the use of self-monitoring tools among older adults reported that data taken from wellness monitoring tools could be used as an education, tracking, and problem solving tool as well as an indicator on how to prioritize care12. The same study interviewed community-dwelling older adults and found positive reactions to the idea of sharing information from wellness trackers with their doctors. Both health care providers and older adults believed that sharing wellness information would improve patient-doctor communication; however, researchers found that older adults in the study were not likely to adopt self-monitoring tools for everyday use due to low perceived personal usefulness and control over data privacy12. Even when this technology is found acceptable, patients’ and providers’ expectations for use of the data in healthcare do not necessarily align13.

Wearable activity trackers can track step count, distance walked, and calories burned. They also have the capability of setting goals, posting to social media sites, creating networks with friends and family, and displaying visual presentations of data if the user syncs their device with a smartphone or computer. For these devices to be effective in motivating health behavior change, they need to be affordable, accurate, and comfortable to wear. The data collected needs to be displayed in a way that is easy to access and interpret by the user14. If these needs are not met, despite good intention, these devices may pose substantial barriers to adoption and use among older adults who, on average, are not as comfortable with technological devices as younger adults15.
Studies that have assessed wearable activity tracker usability in older adult populations and in populations with chronic conditions have yielded mixed results. In a study of older adults with chronic disease, researchers found that participants thought wearable fitness trackers were comfortable and increased awareness of their physical activity, but lack of instructions and limited outside assistance on how to use the tracker were barriers to adoption. Another study assessing the usability and validity of the Fitbit among patients with chronic obstructive pulmonary disease (COPD) found that participants gave the Fitbit high usability scores and reported that the device was pleasant and easy to wear. In a small study of older adults who wore three different wearable activity trackers, participants were initially excited about wearing activity trackers, but 5 out of 8 participants stated that they would not continue using the devices because they felt the devices were uncomfortable, inaccurate or a waste of time. Surprisingly, none of the participants stated technological issues as a reason for discontinuation of use. All of these studies concluded that wearable activity trackers could be useful for self-management of chronic disease, but leave open many questions about their acceptability and use in healthcare integration.

Given the enormous potential for wearable fitness trackers to improve the physical health and well-being of patients with chronic diseases, it is important to further characterize and address any barriers to use and assess the feasibility of adoption of these technologies among older populations. We conducted an exploratory investigation using a 3-week field study with qualitative feedback to capture acceptability of using wearable activity trackers and attitudes towards integrating the use of these devices into clinical care among men with prostate cancer.

Methods

We undertook the Physical Activity and Sedentary Time (PAST) project to better understand device-measured physical activity and sedentary time patterns in men with prostate cancer including those with a history of ADT use. One piece of the project we report on here is a 3-week field study using mixed methods to better understand prostate cancer patients’ acceptability of fitness tracking and its potential to be used within health care. We divided the field study in two phases: an initial one-week baseline phase with required use of one commercially available fitness tracker (Fitbit Zip) followed by a 2-week phase of optional use. We conducted qualitative in-depth interviews at the end of the 3-week study.

Study Population and Recruitment

Men with prostate cancer (N=31) were recruited from a medium sized health care system in Washington State from October 2014 to May 2015. Human subjects approval was obtained from the Group Health Research Institute. Potentially eligible participants were identified using electronic health record data. Men with Surveillance, Epidemiology, and End Results (SEER) codes indicating prostate cancer without distant or metastasized disease were included (summary stages 0-5). We excluded men with encounter codes for palliative care, or diagnosis codes indicating a serious mental health disorder or substance use disorder. Letters were mailed to eligible men inviting them to phone a study staff member if they were interested in participating. On the phone, a study staff member completed oral informed consent and asked additional screening questions. Exclusion criteria were not having prostate cancer, unable to stand, unable to walk one block, and not speaking and reading English.

Procedures

Participants attended in-person visit to receive a Fitbit Zip device to wear for one required baseline week and completed a brief baseline survey. The baseline week allowed us to characterize participants’ on step count levels to ensure there was a diverse range of physical activity levels. During the visit we helped participants set up and sync the device to their Fitbit account. Participants were also provided with written instructions on how to wear and sync the Fitbit device. After the baseline phase, participants were given the Fitbit to keep and use for two optional use weeks, along with instructions for how to use and sync the device to a personal computer or smartphone. At the end of this 2-week phase, we conducted an in-depth exit interview by phone. Using an app we built to access data from Fitbit server through Fitbit’s Web Application Programming Interface (API), we downloaded participants’ Fitbit data for analysis of steps. Participants kept their Fitbit and were paid $10 for completing the study.

Data collection and analysis

Collected data included the baseline survey, Fitbit steps, and interviews. The baseline survey collected participant demographics, health characteristics and level of technology use and ownership from items used in the National Health and Aging Trends Survey. We analyzed surveys with descriptive statistics to characterize participants and compare participant groups by those who had received ADT and those who had not.
To describe physical activity levels, average daily step counts were computed for each participant during the baseline phase and for those who used the Fitbit in the optional use phase. We noted whether or not participants had synced their Fitbit on their own during the optional use phase using the API (indicated by having step count data on at least one day during the optional use phase and coded as yes/no). We also asked the participant whether they used their Fitbit (yes/no) in the optional use phase during the exit interview. We classified prior use of a wearable physical activity tracker (e.g. pedometer, Fitbit, Jawbone) before participating in the PAST study (yes/no) from answers to the baseline survey or during the exit interview.

Qualitative data was collected through semi-structured in-depth interviews to capture attitudes on acceptability of the Fitbit for personal use and interest in sharing Fitbit data with health care providers through the electronic health record. Interview questions were open ended with follow-up prompts. Example questions included:

- What did you think of the Fitbit? What did you like about wearing the Fitbit? What didn’t you like about wearing the Fitbit?
- How have you used your Fitbit during the past 2 weeks? What would have helped you use the Fitbit more?
- Would you want a healthcare team member to see your Fitbit data? If so, who?
- Would you be open to having your Fitbit data go into your electronic medical record? What concerns would you have about this?

The interviews were audio-recorded and transcribed. A coding team of 4 members reviewed transcripts, developed codes and definitions (i.e., codebook), and refined the codebook in an iterative process using inductive thematic analysis. Transcripts were coded using Atlas.ti software to assist with summarizing quotes for each code and identifying themes.

**Results**

**Participants**

Out of 205 people contacted for participation, 31 (15%) completed the in-person visit. One participant dropped out of the study, and two participants were unable to complete exit-interviews. Two participants did not wear their Fitbit Zip properly during the baseline phase and were not included in the analysis. Of the 26 participants who completed the study, 14 (54%) had ADT treatment. Men with a history of taking ADT were older, were more likely to be retired, had lower body mass index (BMI) and had a lower daily step count (over 2000 fewer steps on average) than men without a history of ADT use (Table 1). Men in both groups had many chronic conditions and were relatively similar in their use of technology and ownership of various types of devices. Only about half of the men used email and texting regularly, but all owned a cell phone. Only six participants reported owning fitness trackers and these were primarily men with no history of ADT treatment.

**Physical Activity**

Table 2 shows participants’ Fitbit use by group. During the baseline week, participants wore the Fitbit for a minimum of 5 days. Daily steps ranged from 2041 to 11205. Fourteen participants (54%) reported using the Fitbit during the optional use two-week phase. Of these, six (43%) had received ADT and seven (50%) had never used a pedometer or wearable activity tracker before participating in this study. The data in Table 2 indicate that many men taking ADT, and several with very low step counts (< 4,000 steps/day) were able and willing to use a new technology for fitness tracking during the optional use phase. Of the 8 men without a history of using ADT who used the Fitbit during the optional use phase, all but two men had prior experience using a fitness tracker and baseline phase step counts were a mixture of relatively low (~5,000 steps/day) to high (~10,000 steps/day).

**Attitudes Toward Use**

Themes that emerged from exit interviews regarding attitudes towards Fitbit use include wearability, ease of using technology, value in use, barriers to use, helpful features, and integration with healthcare. Next we describe each theme and provide illustrative quotes.
Wearability

Most participants (93%) found the Fitbit Zip easy to wear and comfortable. Participants noted that the device was easy to put on and said they forgot that they had it on and that wearing the device did not inconvenience them in any way:

*Once I put it on in the morning I was totally unaware of its presence on my body or in my pocket.* P17

Very few problems wearing the Fitbit were reported. A few participants expressed that they sometimes forgot to put on the device in the morning or to reattach the Fitbit when they changed clothes:

*Sometimes I forgot to take it off, because I was sick and I wasn’t moving around that much. Sometimes I forgot to put it on. I wore it when I remembered.* P14

One participant reported the device fell out of its case and another did not like that they had to remove the Fitbit when they used their hot tub.

---

**Table 1: Participant characteristics and mean Fitbit steps by history of ADT treatment**

<table>
<thead>
<tr>
<th>Demographic characteristics n, (%) unless otherwise specified</th>
<th>Total n = 26†</th>
<th>History of ADT n = 14†</th>
<th>No ADT treatment n = 12†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years, mean (SD)</td>
<td>70.5 (9.7)</td>
<td>74.4 (7.9)</td>
<td>65.8 (9.9)</td>
</tr>
<tr>
<td>Retired</td>
<td>18 (69.2)</td>
<td>13 (92.9)</td>
<td>5 (41.7)</td>
</tr>
<tr>
<td>Married</td>
<td>19 (73.1)</td>
<td>12 (85.7)</td>
<td>7 (58.3)</td>
</tr>
<tr>
<td>Some college or less</td>
<td>5 (19.2)</td>
<td>4 (28.6)</td>
<td>1 (8.3)</td>
</tr>
<tr>
<td>Non-Hispanic white</td>
<td>21 (80.8)</td>
<td>12 (85.7)</td>
<td>9 (75.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Health characteristics</th>
<th>Total n = 26†</th>
<th>History of ADT n = 14†</th>
<th>No ADT treatment n = 12†</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI, kg/m², mean (SD)</td>
<td>28.3 (4.6)</td>
<td>27.0 (3.2)</td>
<td>29.6 (5.6)</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>13 (50.0)</td>
<td>6 (42.9)</td>
<td>7 (58.3)</td>
</tr>
<tr>
<td>High cholesterol</td>
<td>11 (42.3)</td>
<td>5 (35.7)</td>
<td>6 (50.0)</td>
</tr>
<tr>
<td>Arthritis</td>
<td>8 (30.8)</td>
<td>5 (35.7)</td>
<td>3 (25.0)</td>
</tr>
<tr>
<td>Depression</td>
<td>4 (15.4)</td>
<td>2 (14.3)</td>
<td>2 (16.7)</td>
</tr>
<tr>
<td>Years since diagnosis</td>
<td>5.4 (6.7)</td>
<td>9.1 (7.7)</td>
<td>1.3 (0.7)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technology use &amp; ownership</th>
<th>Total n = 26†</th>
<th>History of ADT n = 14†</th>
<th>No ADT treatment n = 12†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer in the home</td>
<td>25 (96.2)</td>
<td>13 (92.9)</td>
<td>12 (100)</td>
</tr>
<tr>
<td>Online in past month</td>
<td>25 (92.6)</td>
<td>13 (92.9)</td>
<td>12 (100)</td>
</tr>
<tr>
<td>Emailed in past month</td>
<td>25 (96.2)</td>
<td>13 (92.9)</td>
<td>12 (100)</td>
</tr>
<tr>
<td>Emailed most days in past month</td>
<td>14 (53.9)</td>
<td>6 (42.9)</td>
<td>8 (66.7)</td>
</tr>
<tr>
<td>Texted most days in past month</td>
<td>8 (30.8)</td>
<td>3 (21.4)</td>
<td>5 (41.7)</td>
</tr>
<tr>
<td>Accessed internet on mobile device in last month</td>
<td>18 (69.2)</td>
<td>9 (64.3)</td>
<td>9 (75.0)</td>
</tr>
<tr>
<td>Own a tablet</td>
<td>13 (50.0)</td>
<td>7 (50.0)</td>
<td>6 (50.0)</td>
</tr>
<tr>
<td>Own a laptop</td>
<td>17 (65.4)</td>
<td>8 (57.1)</td>
<td>9 (75.0)</td>
</tr>
<tr>
<td>Own a smart phone</td>
<td>12 (46.2)</td>
<td>5 (35.7)</td>
<td>7 (58.3)</td>
</tr>
<tr>
<td>Own a cell phone</td>
<td>14 (53.9)</td>
<td>11 (78.6)</td>
<td>3 (25.0)</td>
</tr>
<tr>
<td>Own a fitness tracker</td>
<td>6 (23.1)</td>
<td>1 (7.1)</td>
<td>5 (41.7)</td>
</tr>
</tbody>
</table>

| Fitbit steps per /day, mean (SD)                              | 6,239 (2,564) | 5,139 (2,114) | 7,521 (2,517) |

ADT = androgen deprivation therapy
† Numbers may not add up to totals due to missing values

---

**Wearability**

Most participants (93%) found the Fitbit Zip easy to wear and comfortable. Participants noted that the device was easy to put on and said they forgot that they had it on and that wearing the device did not inconvenience them in any way:

*Once I put it on in the morning I was totally unaware of its presence on my body or in my pocket.* P17

Very few problems wearing the Fitbit were reported. A few participants expressed that they sometimes forgot to put on the device in the morning or to reattach the Fitbit when they changed clothes:

*Sometimes I forgot to take it off, because I was sick and I wasn’t moving around that much. Sometimes I forgot to put it on. I wore it when I remembered.* P14

One participant reported the device fell out of its case and another did not like that they had to remove the Fitbit when they used their hot tub.
Ease of using technology

Six participants successfully used a smartphone or computer to sync their Fitbit during the optional use phase, although the syncing process was challenging for some of these participants. A few men reported needing help from a family member to successfully sync their devices:

*My wife is whiz kid on the computer and so she would call me in at the end of each day just before we went to bed and say, “Okay, stand here so the computer can read what’s on your Fitbit ... sometimes I had to stand a little – there a few minutes before the information was transferred from my Fitbit to the computer, but then it all lit up in color and it was, not a bar graph, but whatever you call that, a graph. And that was very interesting. It was easy to read, easy to understand.* P17

However, four participants found it easy to sync their devices and considered the process user friendly and presented data they found interesting, for example one participant reported:

*Every day or two I sync it. I download it to Fitbit application for my iPhone and so every day or two I sync it. And then I just sort of look at the information there ... It’s very easy. The Fitbit application is very – I think the term is user friendly.* P5

The remaining three participants had difficulty syncing their Fitbits by themselves:

*I guess that’s one of the things about this particular FitBit that was somewhat more difficult because I had to go online to the FitBit webpage to find out why I couldn’t get it to sync; it did not sync in the beginning. So I had to do a number of things to make it work.* P24

For some participants, the frustration led to giving up on using the device:

---

Table 2: Fitbit steps and use by treatment group

<table>
<thead>
<tr>
<th>ID</th>
<th>Baseline steps</th>
<th>Used Fitbita</th>
<th>Synced Fitbitb</th>
<th>Prior use of trackerc</th>
<th>ID</th>
<th>Baseline steps</th>
<th>Used Fitbita</th>
<th>synced Fitbitb</th>
<th>Prior use of trackerc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>15</td>
<td>6361</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>7916</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>16</td>
<td>5108</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>3370</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>17</td>
<td>5002</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>4495</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>18</td>
<td>6648</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>8696</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>19</td>
<td>10138</td>
<td>No</td>
<td>Tried*</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>3645</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>20</td>
<td>4063</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>4553</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>21</td>
<td>9081</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>2041</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>22</td>
<td>7324</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>5897</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>23</td>
<td>11013</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>6092</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>24</td>
<td>5149</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>7841</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>25</td>
<td>9167</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>12</td>
<td>2832</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>26</td>
<td>11205</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>3988</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3491</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Count:</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td></td>
<td>Total Count:</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

*a Could not sync device due to technical difficulties so did not use the Fitbit
*b Reported using the Fitbit during the 2-week period of optional use during the interview
*c Based on accessing the Fitbit API to determine whether or not they synced their Fitbit during the optional use phase

---

1054
Largely, I’m not wearing it because it doesn’t interact with my computer very easily...why bother? I just go use my manual step counter. P11

Participants who did not sync their Fitbit said they did not try because they were either too busy or did not care about connecting to their account because they could use the display on the device without using their computer or smartphone. In other cases, participants attempted to sync their device but encountered technical difficulties they could not overcome. Indeed, difficulties using technology were a common reason participants gave for not using the Fitbit during the optional use phase.

Value in using

Participants were divided in their opinions on the usefulness of the Fitbit for increasing their activity levels. For some, the Fitbit improved their awareness of their own physical activity level and motivated them to increase it by setting daily step count goals or helped them maintain their activity level:

I think I like to make sure I’m doing some minimal amount of activity. And it’s kind of fun to see what you’ve been doing, how many steps you’ve done, how many miles you’ve gone. P23

Well, I think it was fascinating to find out how many steps I had done for one thing or another and it was helpful on the walk that my wife and I try to take but fail to take every day. P10

Being able to check their step count throughout the day made them more conscious of how active they had been and enabled them to identify opportunities to increase activity to meet their goal. In other cases, participants reported the device would not be useful because they felt they were active enough and did not need to change their activity levels with or without the Fitbit.

Well, I’m not sure that I need to wear it, because I’m pretty active, like I go golfing, and I go walking, as much as possible, and I go up and down stairs all day long, so I don’t feel like I – I do sit for a long periods of time, also, but I lift weights, and I’m pretty active. P2

Because I’m very active usually and I’m up going and running here and there so I don’t think I need to keep track of it. P3

Participants also liked having the ability to track their own data. Family members were also cited as contributing to the usefulness and motivation aspects of the Fitbit if they had similar devices and could serve as challengers in activity competitions. Participants also felt the Fitbit motivated them by providing a reward:

So I think it builds in a positive feedback loop that it can provide, and that little instrument, the FitBit, is a really powerful way of doing that, so I think it – I have mostly all positive things to say. I don’t think there’s any negative, maybe you just have to remember to log in on some point, sync your FitBit to your computer; that’s about the only thing that you have to do. P24

People like to get...rewards, and by being able to look at it easily online and on my phone. I know both my wife and I have sometimes been at 10:00 p.m. occasionally dressed for bed and see that we only need, oh, 600 more steps to get to that 10,000 and go put your coat on and walk around the block. P22

Many participants endorsed value in continuing to use the Fitbit and expected they would continue to use the device indefinitely.

Barriers to Fitbit use

Conversely, some participants reported that the Fitbit was not helpful because they felt unable to alter their level of physical activity due to physical health challenges like pain and injuries. Fatigue and poor stamina limited participants’ ability to walk for long periods or to engage in other activities.

My situation is if I did not have pain, I would be active enough that I wouldn’t need the Fitbit because I would know I’m getting – I’m not very sedentary – well, I’m retired, so I guess I am sedentary. But when I don’t have pain and I’m not limited by the energy that I have from my treatment, I’m quite active. P1
Another barrier included several participants’ concern that the step count is inaccurate or did not capture important activities of daily living in which other forms of physical activity are used (e.g., muscle-strengthening, stretching):

So I'll give you a case. I filled my laundry, and it's logged I walked 2,000 steps. I did not walk 2,000 steps. P20

I guess the only surprise was that it seemed to register less activity than I felt I actually did because it was only measuring steps, and I was doing more than steps. I was lifting. I was bending. I was twisting. I was doing all that other sort of stuff. P11

Priority Fitbit features

Step count and distance walked were the two features used most by participants. Some participants used social features with family members. Very few participants reported using other Fitbit features (e.g., calories, active minutes, challenges):

I’d only use it [Fitbit] for a pedometer. I mean that’s all. I don’t give a [darn] about the calories and all that stuff.” P4

I think I like to make sure I’m doing some minimal amount of activity. And it’s kind of fun to see what you’ve been doing, how many steps you’ve done, how many miles you’ve gone. P23

Two participants who had continued to wear the Fitbit during the optional use phase and had also previously owned Fitbits reported using the social features:

I like it. I mean, it gives me weekly updates. Every now and then we’ll challenge our daughters because the whole family has one now. We bought them for them, too. So we’ll do a challenge every now and then, and I'll try to kick butt. P21

I've got some family members that are "Fitbit friends," so we kinda see each week how many steps everybody walked. P20

Attitudes toward integrating Fitbit with care

When asked about whether they were open to having their Fitbit data go into their electronic health record, participants felt comfortable sharing their Fitbit data for this purpose though some endorsed the need for limitations:

Yeah. I think it keeps people honest in that case. But at the same time, I don’t wanna start getting ads for granola bars every day in my inbox. P18

That would be fine with me. I think if it would help a physician or someone understands how you’re doing, there would be no problem with that. P23

Other participants describe additional information that would be needed to contextualize data shared with providers:

That’s a good question because, I think I most generally would. I think there might need to be some notes along with it, like I said, if there was a decrease in activity it might be because of illness, or something like that. So I think it tells part of a story, but not all of a story. If you did that, you would need to have...something like, a mechanism usually about the person’s physical ability to walk or whatever. P24

Many felt it would be helpful information for their provider to have to ensure their health is supported and that interventions are provided at appropriate times:

Again, somebody I had a relationship with, no matter who it was as long as it was where you had already established something and could be looking at that and either saying, “Well, that was just great. It was just 10,011 steps yesterday.” Or, “I saw you were only doing 1,000 steps. Are you feeling okay? Is there something getting in the way of your walking?” That kind of stuff. P10

Clearly, the primary care physician, the one that’s supposed to be your first point of contact, and so the ongoing collaborator in your wellbeing would be essential. And then depending on each person’s situation, and the teams wrapped around, and whatever team would seem appropriate to be looking at that data that would help them design a better intervention or better and better treatment that all makes sense to me. P18
Other participants were not sure how it would be helpful to their health care team and worried it might burden physicians, in particular:

*The thing for me would be, this would be all right for them, but if it’s just one little piece of information going into my record, I don’t see it making much difference one way or the other. And it’s the kind of thing that I don’t, and I may be wrong, I don’t see doctors or even nurses spending a whole lot of time looking at my Fitbit data.* P19

Many participants suggested that someone other than the physician such as a physical therapist or personal trainer should be the person to go over their data with them for feedback.

**Discussion**

Findings from our 3-week exploratory field study suggest that some men with prostate cancer find wearable fitness trackers like Fitbits highly acceptable—participants generally found the device comfortable and easy to wear, averaging over 6,000 steps per day when used. When given the choice, over half of participants continued using the Fitbit during the optional use two-week phase. While participants found step count and social features most useful, they experienced several barriers to use, including health-related limitations, problems syncing devices, and data inaccuracies and omissions. Despite these issues, many participants expressed value in sharing Fitbit data with their health care team. These insights expand on prior work by further characterizing feasibility of wearable fitness tracker adoption among an older population, men with prostate cancer.

Our qualitative themes align well within existing technology acceptance frameworks (e.g. the Technology Acceptance Model) and prior work. The TAM emphasizes the importance of perceived usefulness and perceived ease of use in shaping actual system use. We demonstrated actual use by having men use a Fitbit during a baseline week and two optional use weeks. While men were very willing to wear the device for the required baseline week, only half of the men used the Fitbit during the optional use phase that followed. Men with a history of ADT treatment were less likely to use the device than men without an ADT treatment history during the optional use phase. One factor that likely contributed to this disparity is the difference in average age between these two groups: the average age of ADT users in this sample was 10 years higher than those who had not received ADT. Also, men with a history of ADT use had less prior experience using fitness trackers. However, it is promising that many ADT users with very low step counts and who did not have prior experience using such fitness tracking technologies continued using their Fitbit during the optional use phase.

Regarding perceived ease of use, participants generally found the Fitbit to be easy to use and comfortable. However, many experienced technological barriers when trying to sync their Fitbit, which discouraged them from using the device. Some gave up trying to use their Fitbit while others sought help from family members that were more experienced with using technology. We selected the Fitbit Zip because it does not require smart phone ownership (which was low in our sample; 56%) and participants could sync it to a computer (96% owned a computer) or view the step count and distance walked for the current day on the device display. Therefore, simple tracking devices appear preferred by the older men in this study.

One design recommendation that could help improve ease of use includes continuing to offer fitness trackers that have a display that do not require syncing. Furthermore, having a medical team member request that patients use a fitness tracker and work with them over time to understand the data and learn to use more technologically advanced features of devices (e.g. syncing) could help promote a positive feedback loop that many participants were excited about. Giving patients clear and simple instructions for using wearable tracker devices, syncing them, and using their features could also facilitate adoption. This recommendation dovetails well with prior work indicating that health care providers could work together with older adults to help them understand their wellness monitoring data.

Regarding perceived usefulness, many participants endorsed value in using the Fitbit to ensure they engaged in a “minimal amount of activity” (P23) and to motivate them to squeeze in more steps. Barriers to usefulness included health limitations and feeling very active already. The latter is somewhat concerning considering that the average step count was well below recommendations for older adults or those with chronic conditions (8,000 steps/day). Particularly among men with a history of ADT use, step counts were very low (~5,000 steps/day). This further underscores the importance of helping men with prostate cancer receive feedback and education on their level of physical activity and its importance in supporting them as they age. Many men were quite active to start with, although wearable trackers like Fitbit could help these men continue to stay active as levels of physical activity decline with age.
All of the men were willing to wear the Fitbit if asked by someone from their health care system. This suggests that if used in clinical practice to get a sense of habitual activity, a health care provider might be able to discuss the importance of being even more active or continue to remain active. Indeed, all men reported willingness to share their Fitbit data with their health care provider and few had concerns about the information becoming part of the electronic health record. This willingness could, in part, be due to their involvement in a study that took place within their own health care system. There were some limitations on willingness to share the data including not wanting to be bombarded with messages about their health and being able to amend the data in order to explain or contextualize it.

Much prior work that has examined physical activity trackers in older adults has been limited to perspectives on use or required use over a week or less. Our 3-week field study enabled us to capture their use for longer. When offered the opportunity, many participants opted to continue using the Fitbit during the optional two-week phase but faced several technical barriers. While this method offered further insights into adoption, our small sample of men with prostate cancer and other chronic conditions could limit the generalizability of findings to other groups or devices. Additional support for using and syncing Fitbit devices could have impacted participants’ overall experience or improved optional use. There is merit in future work to further characterize and address the physical activity tracking needs of older adults and connection with healthcare providers. In particular, studies are needed to examine longer-term use of devices in larger and more diverse samples of older adults. Our findings suggest such devices should be designed to be simple for older adults to use, incorporate social features, and capture additional physical activities beyond walking, such as muscle strengthening and stretching. More targeted and easy to technology could improve adoption by older adults. Future work is also needed to explore how to integrate physical activity data into clinical care, including information and workflow needs of health care providers, in ways that minimize burden and promote utility. It may not be feasible to fully review and contextualize a detailed data stream within short provider visits. Future work should examine the right places for this data to enter into clinical workflow and the right health care team members to receive, review, and provide patient feedback on the data.

Conclusion

Our exploratory field study found that there are many opportunities to link fitness tracking to clinical care and to help the broader population of men with prostate cancer, and possibly other chronic health conditions, become more active and prevent further health declines. Our findings are congruent with those of Huh’s work in people with various chronic conditions and Vooijs’s work with COPD. This accumulation of evidence suggests that small-scale feasibility studies integrating fitness trackers with clinical care for high risk populations (e.g. COPD, prostate cancer) are warranted. These future interventions will need to make the technology as straightforward as possible, most likely through in-person demonstration sessions in conjunction with simple written materials and interactions with the health care team.

References

GT2RDF: Semantic Representation of Genetic Testing Data

Anamika Paul Rupa, MS, Sweta Singh, Qian Zhu, PhD
Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD, US

Abstract

Accelerated by the Human Genome Project, genetic testing has become an increasingly integral component in diagnosis, treatment, management, and prevention of numerous diseases and conditions. More than 480 laboratories perform genetic tests for more than 4,600 rare and common medical conditions. These tests can effectively help health professionals to determine or predict the genetic conditions of their patients. However, physicians have not actively incorporated such innovative genetic technology into their clinical practices according to two national wide surveys commissioned by UnitedHealth Group. To fill the gap of insufficient use of a large number of genetic tests, we generated a single Resource Description Framework (RDF) resource, called GT2RDF (Genetic Testing data to RDF) by integrating information about disease, gene, phenotype, genetic test, and drug from multiple sources including Genetic Testing Registry (GTR), Online Mendelian Inheritance in Man (OMIM), MedGen, Human Phenotype Ontology (HPO), ClinVar, National Drug File Reference Terminology (NDF-RT).

Meanwhile, we manually annotated and extracted information from 200 randomly selected GeneReviews chapters, and integrated into the GT2RDF. We performed two case studies to demonstrate the usability of the GT2RDF. GT2RDF will serve as a data foundation to support the design of a genetic testing recommendation system, called iGenetics, which will ultimately facilitate the pace of precision medicine by means of actively and effectively incorporating innovative genetic technology in clinical settings.


Introduction

30 million people in the United States are living with rare diseases, and 80% of rare diseases are genetic. According to a recent survey of patients, family members, physicians and allied health care professionals regarding physician training in rare disease, these patients visit an average of 7.3 physicians before receiving an accurate diagnosis. Survey results indicated that both primary care physicians and specialists were equally likely to consult someone they considered a “disease expert” to help them make the diagnosis, which can potentially lead to the treatment delay. Consistently, forty-four percent of patients agreed with the statement, “Because of a slow diagnosis, treatment was delayed and the impact on my condition has been negative.”

This survey concludes that there is an urgent need to bridge the knowledge gap by educating and connecting patients, families, physicians and specialists, and urges families to educate themselves and their doctors about genetic testing technologies that may expedite an accurate diagnosis and lead to effective timely and appropriate treatment.

With the advance in genetic technology, more than 680 laboratories perform genetic tests for more than 4,600 rare and common medical conditions. No wonder so much hope is riding on the promise of “precision medicine”, in which genetic screening and other tests give doctors more evidence for tailoring treatments to patients, potentially improving health care and saving money. However, physicians have not actively incorporated such innovative technology into their clinical practices according to two national wide surveys commissioned by UnitedHealth Group in conjunction with Harris Interactive (n=2,760). About half of the physicians surveyed stated that the lack of familiarity with genetic tests is the main barrier to apply them in their regular practice, and over three-quarters of physicians are either somewhat or very concerned about the lack of evidence supporting the use of genetic testing. Meanwhile, these physicians identified four possible solutions that would possibly encourage them actively use of these genetic tests, a) enhanced medical and continuing education, b) provision of clinical decision support guidance, c) better access to clinical evidence, and d) affordable tests. To meet physicians’ needs, in this study, we designed and generated a centralized data resource about genetic testing in RDF, in order to provide a comprehensive view of integrative genetic testing information and recommend appropriate genetic tests and relevant information to support clinical decision-making.
A large volume of genetic disease related information has been accumulated, as a result of efforts and collaborations made by the NIH and other stakeholders. These include the primary resources, GTR(6), ClinVar(7), MedGen(8), OMIM(9), HPO(10), NDF-RT(11). In addition, genetic guidelines are continuously published for public use. For instance, GeneReviews(12), pharmacogenomics guidelines published by the Clinical Pharmacogenetics Implementation Consortium (CPIC)(13), Evaluation of Genomic Applications in Practice and Prevention (EGAPP)(14), a systematic, evidence-based process for assessing genetic tests and other applications of genomic technology. Although these resources provide invaluable information about genetic tests, distributed storage and diverse data representation lack interoperability and capability for its better integration and adoption at an institutional level. The effort required for adopting, ensuring appropriate protections, integrating, and maintaining such information within clinical infrastructures remains significant. Semantic web technology (SWT) is used to construct a common framework that accelerates data sharing and reuse across different applications, institutions, and community boundaries. SWT proposes to apply RDF as a flexible data model, which “effectively capture data semantics and enable semantic query and matching, as well as efficient data integration.”(15) A variety of RDF-based Semantic Web resources have been generated for biomedical data. For instance, Bio2RDF(16) offers a strategy to generate and query RDF based biological data with around 4 billion RDF triples across over 30 biological resources. Chem2Bio2RDF(17) aggregates data from multiple chemogenomics repositories in RDF. However, to our knowledge, no effort has been made for generating RDF based genetic testing data to support future clinical decision-making. In this present work, we generated an RDF resource, GT2RDF by integrating genetic testing relevant information from multiple well-known data resources, including GTR, OMIM and GeneReviews and incorporating a meta-ontology defined in OWL (Web Ontology Language) to formally capture the semantics of GT2RDF. We present use cases of how GT2RDF can be applied in seeking pertinent information regarding genetic tests and genetic diseases via SPARQL Protocol and RDF Query Language (SPARQL).

In the rest of paper, we briefly introduce the materials used in this study in the Materials section. We describe details about the generation of GT2RDF in the Methods section. Then we present the result and case study in the Result section. Finally, we discuss the benefit we obtained, the lesson we learnt from this study and our future direction in the Discussion section.

Materials

GT2RDF integrates diverse types of information extracted from seven different resources, which are briefly described below.

GTR(6) provides a central location for voluntary submission of genetic test information by providers. We downloaded GTR by November 1, 2015. We primarily accessed ‘test_condition_gene’ and ‘test_version’ files from GTR to retrieve information about test, disease and genes.

ClinVar(7) is a public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. We extracted disease, gene and allele information by accessing ‘gene_condition_source_id’ and ‘variant_summary’ files by November 1, 2015.

MedGen(8) organizes information related to human medical genetics, such as attributes of conditions with a genetic contribution. We accessed ‘MedGen_HPO_OMIM_Mapping’ file by November 1, 2015, that contains mappings among MedGen, HPO and OMIM. We extracted disease and phenotype related information and association between disease and phenotype from MedGen.

OMIM(9) is a comprehensive, authoritative compendium of human genes and genetic phenotypes with the primary focus on the relationship between phenotype and genotype. We extracted Disease and Gene information and the association between them by accessing ‘morbidmap’ file from OMIM by November 1, 2015.

HPO(10) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. We extracted Gene and Phenotype information from ‘diseases_to_genes_to_phenotypes’ file by November 1, 2015.

GeneReviews(12) is expert-authored, peer-reviewed disease descriptions “chapters” presented in free text and focused on clinically relevant and medically actionable information on the diagnosis, management, and genetic counseling of patients and families with specific inherited conditions. We accessed GeneReviews by September 10, 2015, and extracted gene, phenotypes, genetic tests and family risk information associated with particular genetic diseases.
**NDF-RT** is a collection of drug relevant information and it provides RESTful web services to access those data. We collected drug and disease information by specifying particular parameters, ‘may_treat’, ‘has_contradiction’ from NDF-RT via its REST API by February 24, 2016.

**Methods**

In this study, we generated an integrative and standardized data resource for genetic test via multiple steps, 1) we semi-automatically collected genetic testing relevant data from multiple resources; 2) we normalized the collected data in standard form for further data integration; 3) we designed a meta-ontology using Protégé in order to capture the semantic meaning and semantic relations presenting in the collected data; 4) we generated and published GT2RDF by incorporating the meta-ontology with the implementation of D2RQ. The workflow performed in this study is shown in Figure 1.

![General workflow of GT2RDF](image)

**Figure 1. General workflow of GT2RDF**

1. **Data Collection.**

The seven data resources mentioned in the Materials section vary in their data representation forms. Thus, we applied different strategies to collect them accordingly.

**Structured data collection.** OMIM, GTR, HPO, MedGen, ClinVar are in structured form. Once we downloaded the data from their websites, we uploaded them into a MySQL database directly. We extracted drug and disease information from NDF-RT by invoking its REST API via two steps, 1) retrieving all disease and drug information by passing ‘DISEASE_KIND’ and ‘DRUG_KIND’ as an input parameter respectively to invoke one API resource named “allconcepts” which returns concept information for specified kinds; 2) disambiguating relations between drug and disease, i.e., indication and contraindication. We extracted indications and contraindications for each drug by filtering “rolenames” by “may_treat” and “CI_with” respectively from the
returned drug information by invoking the API resource named “allInfo” which returns concept information for a specified concept.

**Unstructured data collection.** GeneReviews provides tremendous information about significant patient characteristics relevant to individual genetic diseases, which are critical for further genetic test prediction. In this study, we (SS) manually annotated information presenting in the sections of Diagnosis, Clinical Characteristics, Management and Genetic Counseling for 200 GeneReviews Chapters. In the section of Genetic Counseling, it contains detailed information regarding genetic risk assessment, carrier detection, prenatal testing, and related genetic counseling issues for family members. In this study, we primarily focused on “risk to family members”, which includes information about the parents, sibs, and offspring risk of being affected, being a carrier or being unaffected. Given the complex structure of the section of Genetic Counseling that is out of the scope of this study, we only included ones with explicit content information including the risks of family members. In addition, we did not incorporate information presenting in the sections of Treatment and Prevention although we extracted them, as more information extraction process is needed to retrieve more refined annotation to be integrated into the GT2RDF, which will be proposed as the future work. AR and QZ have manually evaluated the annotation results and corrected the errors and problems. The corrected annotations have been integrated into the database.

2. **Data Normalization.**

To facilitate data manipulation and integration, data normalization is a key. In this study, while we mapped the collected concepts to Unified Medical Language System (UMLS) (21) by using MetaMap (22), we also maintained the existing identifiers as standard representation, such as Test ID from the GTR as an identifier for genetic test; HPO ID as an identifier for Phenotypes; OMIM number as an identifier for genetic diseases; rsID as an identifier for SNP; gene symbol for genes; NUI as NDF-RT unique identifier for drugs. More specifically, we automatically mapped a unique list of the diseases, phenotypes, genetic tests, and drugs to UMLS via the batch process function available at MetaMap. The mapping results generated by the MetaMap, include the UMLS preferred terms along with mapping scores. In order to exclude false positive annotations and labor insensitively manual evaluation, in this study, we only included the annotations with mapping score equaling to “1000”, which refers to an exact mapping between original concept and the preferred terms from UMLS. As only 200 GeneReviews chapters have been manually processed, the number of extractions is manageable, so we manually selected annotation results generated by the MetaMap, in order to maintain the high mapping accuracy. Due to some non-standard terms are used among the aforementioned resources, missing mappings were observed even the “1000” score has been applied. We manually evaluated a random set of annotation results to ensure the accuracy of mapping prior to the GT2RDF generation, in addition, we analyzed the causes of missing mappings for further extension.

3. **Meta-ontology Design.**

In order to capture the semantic meaning and semantic relationships among the collected concepts, e.g., diseases, gene, allele, genetic tests, phenotypes, risks of family members, drugs, we designed a meta-ontology using Protégé (16).


**OWL Object Property Definition.** Object properties are defined in a way to represent the relationships among classes. The list of Object properties is shown in table 1.

<table>
<thead>
<tr>
<th>Object Property</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssociatedWithPhenotype</td>
<td>Domain: ‘Disease’ and Range: ‘Phenotype’</td>
</tr>
<tr>
<td>AssociatedWithGene</td>
<td>Domain: ‘Disease’ and Range: ‘Gene’</td>
</tr>
<tr>
<td>TestForDisease</td>
<td>Domain: ‘Disease’ and Range: ‘Test’</td>
</tr>
<tr>
<td>MayTreatedBy</td>
<td>Domain: ‘Disease’ and Range: ‘Drug’</td>
</tr>
<tr>
<td>HasContraindicationWith</td>
<td>Domain: ‘Drug’ and Range: ‘Disease’</td>
</tr>
</tbody>
</table>

Table 1. Description of Object Properties
**OWL Data Property Definition.** Data properties link individuals to data values. The defined data properties for each class, along with some explanation are shown in Table 2. It is worthy to note that we included information from NDF-RT, which contains non-genetic diseases, thus we defined an additional data property, ‘IsGenetic’ for “Disease” to distinguish genetic diseases from non-genetic diseases.

<table>
<thead>
<tr>
<th>Data Property</th>
<th>Related Class</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiseaseMIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrugName, UMLS_CUI, NUI</td>
<td>Drug</td>
<td>NUI: NDF-RT Unique identifier.</td>
</tr>
<tr>
<td>RiskType, RiskCondition, Relation,</td>
<td>Risk</td>
<td>RiskCondition: Specific condition for the risk, example: “If the parent is affected”. RiskType: The types of risk, Unaffected, Affected or Carrier. Relation: The relationship to the affected patient such as sibs, offspring. RiskPercentage: The percentage under particular risk type with specific “RiskType” for particular family member with a specific “Relation”.</td>
</tr>
<tr>
<td>Relation, RiskPercentage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlleleID, AlleleType, AlleleName,</td>
<td>Allele</td>
<td>AlleleID: Integer value as stored in the AlleleID field in ClinVar. AlleleType: The type of variant. AlleleName: The preferred name for the record. SNP: Reference SNP ID (rs id in dbSNP).</td>
</tr>
<tr>
<td>SNP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhenotypeName, HPO_ID, UMLS_CUI</td>
<td>Phenotype</td>
<td>HPO_ID: Id for HPO.</td>
</tr>
<tr>
<td>TestName, GTR_TestID, UMLS_CUI</td>
<td>Test</td>
<td>GTR_TestID : Lab's test ID (lab_test_id) from GTR.</td>
</tr>
</tbody>
</table>

**4. GT2RDF Generation.**

In this study, we included seven types of information, such as genetic test, disease, genes, genetic variations, phenotypes, medication and family risk according to the definition of the meta-ontology. For the GT2RDF generation, we applied D2RQ, which offers RDF-based access to the content of relational databases without having to replicate it into an RDF store, to transform data in the relational database to RDF. The mapping tool of D2RQ creates a default mapping file by analyzing the schema of an existing database. To map our data with the defined meta-ontology, we manually customized the mapping file accordingly. The data is then published in RDF through the D2RQ server, and can be queried via a D2RQ SPARQL endpoint. We also took an RDF dump from D2RQ into Virtuoso(23) to run federated queries. GT2RDF can be accessed via the URL http://biomedinformatics.is.umbc.edu:2020/.

**Result**

**1. Data collection.**

We collected data from seven data resources to generate the GT2RDF. Table 3 shows the detailed information for the data collection.
2. UMLS Mapping.

To represent GT2RDF in a normal form for further data integration, we mapped the individual terms belonging to the four classes, Disease, Drug, Phenotype and Test to UMLS. It is shown in Table 3 that diseases are extracted from several sources, but only GTR and MedGen provide UMLS mappings. Thus, we mapped the rest of the 15,078 disease terms to UMLS using MetaMap. NDF-RT provides 25,800 UMLS mappings among 26154 drugs and rest of the 354 drug terms have no mappings to UMLS. We tried to map the missing terms in MetaMap and found UMLS CUI for 17 drug names and included them in our database. We mapped all the phenotype terms and test terms using MetaMap because none of the resources provide UMLS mappings for them.

Table 3. Statistical results for data collection.

<table>
<thead>
<tr>
<th>Data types</th>
<th>Number of concepts</th>
<th>Number of relationships</th>
<th>Data resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>34695</td>
<td>745376</td>
<td>ClinVar, OMIM, MedGen, GeneReviews, NDF-RT, GTR</td>
</tr>
<tr>
<td>Gene</td>
<td>32815</td>
<td>702838</td>
<td>GTR, ClinVar, HPO, OMIM, GeneReviews</td>
</tr>
<tr>
<td>Drug</td>
<td>26154</td>
<td>9909</td>
<td>NDF-RT</td>
</tr>
<tr>
<td>Risk</td>
<td>565</td>
<td>565</td>
<td>GeneReviews</td>
</tr>
<tr>
<td>Allele</td>
<td>131708</td>
<td>134970</td>
<td>ClinVar</td>
</tr>
<tr>
<td>Phenotype</td>
<td>8787</td>
<td>108121</td>
<td>HPO, MedGen, GeneReviews</td>
</tr>
<tr>
<td>Test</td>
<td>27624</td>
<td>58905</td>
<td>GTR, GeneReviews</td>
</tr>
</tbody>
</table>

In order to ensure the accuracy of the mapping results, although we only selected UMLS preferred terms with mapping score equaling to 1000, we randomly selected one hundred mapping results for each type of terms and manually reviewed the mapping results. According to our evaluations, there are no incorrect mapping for one hundred Disease and one hundred Drug terms, but there are two incorrect mappings among one hundred Phenotype terms and four incorrect mappings among one hundred Test terms caused due to substring matching. The reasons causing the missing mappings are discussed with examples in the Discussion section. The overall accuracy of UMLS mapping is 98.5% which illustrates the high performance of mapping process using MetaMap. To clarify, “Total Number of Concepts Mapped to UMLS” is part of “Total Number of Concepts” because the concepts with mapping score less than 1000 have not been counted to the mapped concepts. The UMLS mapping result is shown in Table 4.

Table 4. UMLS mapping chart

<table>
<thead>
<tr>
<th>GT2RDF Types</th>
<th>Total Number of Concepts</th>
<th>Total Number of Concepts Mapped to UMLS</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>34,695</td>
<td>28,309</td>
<td>81.6%</td>
</tr>
<tr>
<td>Drug</td>
<td>26,154</td>
<td>25,817</td>
<td>98.7%</td>
</tr>
<tr>
<td>Phenotype</td>
<td>8,787</td>
<td>8,324</td>
<td>94.7%</td>
</tr>
<tr>
<td>Test</td>
<td>27,624</td>
<td>5,498</td>
<td>20.0%</td>
</tr>
</tbody>
</table>

3. GT2RDF generation.

There are total 2,309,014 triples contained in the GT2RDF. Among them 32,815 triples are related to Gene, 34,695 triples are related to Disease, 8,787 triples are related to Test, 26,154 triples are related to Drug, 134,010 triples are related to Allele.

4. Case Study.

In order to demonstrate the usability of the GT2RDF, we performed two case studies accordingly.
Case study 1. Hemophilia A, (24) is a genetic disease caused by missing or defective factor VIII, a clotting protein. Information about diagnosis or management of Hemophilia A is available at OMIM and GTR. However, it is inefficient during the clinical practice to manually review and identify the most significant points. In this case study, we searched against the GT2RDF to retrieve pertinent information about Hemophilia A by composing and executing a SPARQL query shown in the section of “SPARQL query” in Table 5. Specifically, we are searching for related genes, appropriate genetic tests, available drugs, possible clinical features and risk for family members, shown in the section of “Results” in the Table 5. The retrieved risk information shown in Table 5 can be interpreted as “if the proband’s mother is a carrier each male sib is at a 50% risk of having hemophilia A” and “if the proband’s mother is a carrier each female sib is at a 50% risk of being a carrier”.

Table 5. SPARQL query to extract Hemophilia A related information and results

<table>
<thead>
<tr>
<th>SPARQL query</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT distinct ?Disease ?property ?hasValue</td>
<td>Disease Name: HEMOPHILIA A. UMLS CUI: C1366370.</td>
</tr>
<tr>
<td>{?Disease ?property ?hasValue. }</td>
<td>Genetic Test: Factor VIII mutation analysis, DNA Hemophilia A mutation evaluation, DNA Hemophilia A mutation Screen, Molecular Genetic Testing.</td>
</tr>
<tr>
<td></td>
<td>Drugs: 1. antihemophilic factor,</td>
</tr>
<tr>
<td></td>
<td>2. antihemophilic factor, human,</td>
</tr>
<tr>
<td></td>
<td>3. antihemophilic factor, porcine,</td>
</tr>
<tr>
<td></td>
<td>4. factor ix, recombinant,</td>
</tr>
<tr>
<td></td>
<td>5. factor vii, recombinant.</td>
</tr>
<tr>
<td></td>
<td>Phenotypes: Persistent bleeding after trauma, X-linked recessive inheritance etc.</td>
</tr>
<tr>
<td>RiskCondition: If the proband’s mother is a carrier,</td>
<td>RiskCondition: If the proband’s mother is a carrier,</td>
</tr>
<tr>
<td>Relation: Female Sibs, Risk Type: Carrier,</td>
<td>Relation: Male Sibs, Risk Type: Affected,</td>
</tr>
<tr>
<td>Risk Percentage: 50%.</td>
<td>Risk Percentage: 50%.</td>
</tr>
</tbody>
</table>

Table 6. SPARQL query to extract Wilson Disease related information and results

<table>
<thead>
<tr>
<th>SPARQL query</th>
<th>Results</th>
</tr>
</thead>
</table>
In this study, we introduced a novel RDF data resource by integrating heterogeneous information about genetic test from multiple data sources. The GT2RDF serves a data foundation, which will ultimately support the development of clinical decision support (CDS) system, iGenetics. Multiple advantages of the GT2RDF have been observed in the case studies, for instance, it provides comprehensive information regarding genetic disease and genetic test by searching the GT2RDF and other relevant Linked Open Data (LOD) (23). Besides the advantages, we also learnt lesson as well, since we are still in the early stage of this project and much future work is under planning.

Currently, more and more genetic relevant data has been accumulated to support more complex scientific and clinical questions. However, they are represented in their own forms and templates, which hinders its wide use and sharing. In this study, we dedicated our efforts on data normalization with UMLS and other vocabularies, such as OMIM, GTR, and etc., to facilitate further data integration. Showing in Table 4, a majority of selected GT2RDF terms have been mapped to the UMLS via MetaMap, but there is still a small portion of terms have not been mapped to UMLS. The main reason for failed mapping, especially for tests, is because preferred terms with mapping score equaling to 1000 have been selected and we excluded one to many mappings. For instance, one genetic test concept, “ACADM Sequence Analysis”, MetaMap results in two terms, “ACADM” with a score 660 and “Sequence Analysis” with a score 901 for. Similar to drug, “ALIROCUMAB 150MG/ML INJ, SYR, 1ML”, a drug concept with drug signature information like strength, form etc., are mapped to three UMLS terms, “ALIROCUMAB” with a mapping score 645, “ML” with a mapping score 645 and “INJ” with a mapping score 812. In this study, we excluded all of these mappings. But in order to increase the number of UMLS mappings in the future, we will include the mappings with a score less than 1000 and one-to-many mappings (one concept is mapped to multiple UMLS preferred terms). In addition to ensure the mapping accuracy, we will perform manual evaluation afterward. In addition, the primary terminology, UMLS has been applied in the data normalization step, in the future study, we will also utilize other domain vocabularies, such as SNOMED-CT(26) for genetic diseases and LOINC(27) for genetic tests.

Given the demonstrated advantage of SWT in biomedical domain with an increasing number of successful applications, we applied SWT to generate the GT2RDF. Web Ontology Language (OWL) as a web standard language for formally describing the semantics of a set of concepts within a domain, and semantic relationships between those concepts within Semantic Web. In this study, we defined a meta-ontology in OWL to capture the semantics of the GT2RDF. By incorporating with the defined meta-ontology, we formally represented our data in RDF. The meta-ontology and associated annotation not only improves the semantic utility of the GT2RDF, but also decreases the complexity of the SPARQL query (e.g., no need to understand the RDF schema of each data source during the SPARQL construction). Currently, we defined seven primary classes and associated properties in the meta-ontology, which meets our requirement for the current version of the GT2RDF. More complex structure of the meta-ontology will be defined while more data will be integrated into the GT2RDF. In addition, in the experimental phase, we are manually customizing the mapping file for translating original data to RDF based on the meta-ontology, which may not be realistically once we have quite complex data and ontology structures. In the next step, we will move the transformation process in an automated fashion.

By representing data as labeled graphs, RDF provides a powerful framework for expressing and integrating any types of data, and searching across different online resources, which is the outstanding advantage comparing to the relational databases. As of April 2014, under the auspices of an initiative called the Linked Open Data (LOD), more

```

Case Study 2. In this case study, we aimed to obtained more information about the drugs, like dosage level, drug-drug interaction and drug category for Wilson Disease (25) from GT2RDF and DrugBank from Bio2RDF via executing a federated query, which shows the ability to take a query and provide solutions based on information from many different sources. Wilson Disease is an autosomal recessive genetic disorder in which copper accumulates in tissues and it is a treatable disease by using appropriate medications. Thus, more information about drugs used for Wilson disease will be critical for appropriate drug selection. The composed federated SPARQL query and the retrieved results are shown in Table 6.

Discussion and conclusion.

In this study, we introduced a novel RDF data resource by integrating heterogeneous information about genetic test from multiple data sources. The GT2RDF serves a data foundation, which will ultimately support the development of clinical decision support (CDS) system, iGenetics. Multiple advantages of the GT2RDF have been observed in the case studies, for instance, it provides comprehensive information regarding genetic disease and genetic test by searching the GT2RDF and other relevant Linked Open Data (LOD) (23). Besides the advantages, we also learnt lesson as well, since we are still in the early stage of this project and much future work is under planning.

Currently, more and more genetic relevant data has been accumulated to support more complex scientific and clinical questions. However, they are represented in their own forms and templates, which hinders its wide use and sharing. In this study, we dedicated our efforts on data normalization with UMLS and other vocabularies, such as OMIM, GTR, and etc., to facilitate further data integration. Showing in Table 4, a majority of selected GT2RDF terms have been mapped to the UMLS via MetaMap, but there is still a small portion of terms have not been mapped to UMLS. The main reason for failed mapping, especially for tests, is because preferred terms with mapping score equaling to 1000 have been selected and we excluded one to many mappings. For instance, one genetic test concept, “ACADM Sequence Analysis”, MetaMap results in two terms, “ACADM” with a score 660 and “Sequence Analysis” with a score 901 for. Similar to drug, “ALIROCUMAB 150MG/ML INJ, SYR, 1ML”, a drug concept with drug signature information like strength, form etc., are mapped to three UMLS terms, “ALIROCUMAB” with a mapping score 645, “ML” with a mapping score 645 and “INJ” with a mapping score 812. In this study, we excluded all of these mappings. But in order to increase the number of UMLS mappings in the future, we will include the mappings with a score less than 1000 and one-to-many mappings (one concept is mapped to multiple UMLS preferred terms). In addition to ensure the mapping accuracy, we will perform manual evaluation afterward. In addition, the primary terminology, UMLS has been applied in the data normalization step, in the future study, we will also utilize other domain vocabularies, such as SNOMED-CT(26) for genetic diseases and LOINC(27) for genetic tests.

Given the demonstrated advantage of SWT in biomedical domain with an increasing number of successful applications, we applied SWT to generate the GT2RDF. Web Ontology Language (OWL) as a web standard language for formally describing the semantics of a set of concepts within a domain, and semantic relationships between those concepts within Semantic Web. In this study, we defined a meta-ontology in OWL to capture the semantics of the GT2RDF. By incorporating with the defined meta-ontology, we formally represented our data in RDF. The meta-ontology and associated annotation not only improves the semantic utility of the GT2RDF, but also decreases the complexity of the SPARQL query (e.g., no need to understand the RDF schema of each data source during the SPARQL construction). Currently, we defined seven primary classes and associated properties in the meta-ontology, which meets our requirement for the current version of the GT2RDF. More complex structure of the meta-ontology will be defined while more data will be integrated into the GT2RDF. In addition, in the experimental phase, we are manually customizing the mapping file for translating original data to RDF based on the meta-ontology, which may not be realistically once we have quite complex data and ontology structures. In the next step, we will move the transformation process in an automated fashion.

By representing data as labeled graphs, RDF provides a powerful framework for expressing and integrating any types of data, and searching across different online resources, which is the outstanding advantage comparing to the relational databases. As of April 2014, under the auspices of an initiative called the Linked Open Data (LOD), more
than 1000 public datasets from multiple domains (i.e., everything from geographical locations, to mortality data, to species, genes, drugs, and diseases) are available in RDF, and have been integrated by specifying approximately 60 billion RDF triples. LOD has been successfully applied in medical and clinical studies. Federated query allows retrieving more diverse types of information from multiple datasets. In the second case study, we demonstrated the usage of federated query to obtain information related to genetic disease and genetic test, and more information about drugs, such as drug-drug interaction and dosage from the GT2RDF and Linked Open Data (LOD), such as DrugBank. In our future work, we will explore more datasets available from LOD combining with the GT2RDF, in order to provide a more comprehensive view of the interested genetic disease or other queries. Additionally, the communication over the network can quickly become a performance bottleneck, leading to high query latencies, especially for federated query, is running across multiple RDF graphs, which requires transiting data across the network at the time of query in order to match complex SPARQL patterns navigating the RDF graphs. Recently, various solutions for federated query processing of heterogeneous RDF data sources have been discussed in the literature. But none of them were attempting to leverage Big Data techniques, such as MapReduce and NoSQL database to overcome the aforementioned barriers. In our future work, we will explore Big Data techniques to improve the performance of federated query.

In this study, we have integrated multiple resources. However, in order to provide more comprehensive information to support clinical decision-making, we will integrate more data into the GT2RDF. We will primarily focus on several data extensions, 1) GeneReviews has been widely used in the clinical settings, physicians are most likely to seek solutions from GeneReviews by manually reviewing the entire chapter, which is time consuming. In the early stage of this project, we manually annotated 200 chapters of the GeneReviews, from where we successfully retrieved information about genes, diseases, phenotypes, treatments, and family risk from multiple sections. In the next step, we will semi-automatically process the entire 649 GeneReviews chapters (at the time of writing) with the existing NLP tools, including MetaMap and NCBO annotator to achieve better annotation results. Meanwhile, we will invite domain experts to review and prove the annotations prior to integrate it into GT2RDF. 2) In order to enhance the scientific strength of the associations presenting in the GT2RDF, we will integrate scientific evidence extracted from scientific literature for each association. To be specific, we will explore Semantic MEDLINE to identify and map relevant literature information, i.e., PubMed ID to each association. 3) Although GeneReviews provides information about clinical characteristics for individual patients, such information is insufficient to support individualized genetic test recommendation. Conversely, Electronic Health Records (EHR) contains a wide spectrum of patient medical information, including billing data, laboratory test results, medication records, clinical documentation and imaging results. In the next step, we will statistically identify significantly clinical characteristics associated with each genetic disease via retrospective study over EHR data and deposit into the GT2RDF for further prediction.

References


Scientific Reproducibility in Biomedical Research: Provenance Metadata Ontology for Semantic Annotation of Study Description

Satya S. Sahoo, PhD1, Joshua Valdez, BS1, Michael Rueschman, MPH2
1Division of Medical Informatics, School of Medicine, Case Western Reserve University, Cleveland, OH; 2Department of Medicine, Brigham and Women’s Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

Abstract
Scientific reproducibility is key to scientific progress as it allows the research community to build on validated results, protect patients from potentially harmful trial drugs derived from incorrect results, and reduce wastage of valuable resources. The National Institutes of Health (NIH) recently published a systematic guideline titled “Rigor and Reproducibility” for supporting reproducible research studies, which has also been accepted by several scientific journals. These journals will require published articles to conform to these new guidelines. Provenance metadata describes the history or origin of data and it has been long used in computer science to capture metadata information for ensuring data quality and supporting scientific reproducibility. In this paper, we describe the development of Provenance for Clinical and healthcare Research (ProvCaRe) framework together with a provenance ontology to support scientific reproducibility by formally modeling a core set of data elements representing details of research study. We extend the PROV Ontology (PROV-O), which has been recommended as the provenance representation model by World Wide Web Consortium (W3C), to represent both: (a) data provenance, and (b) process provenance. We use 124 study variables from 6 clinical research studies from the National Sleep Research Resource (NSRR) to evaluate the coverage of the provenance ontology. NSRR is the largest repository of NIH-funded sleep datasets with 50,000 studies from 36,000 participants. The provenance ontology reuses ontology concepts from existing biomedical ontologies, for example the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), to model the provenance information of research studies. The ProvCaRe framework is being developed as part of the Big Data to Knowledge (BD2K) data provenance project.

Introduction
Scientific results from basic and clinical experiments represent the core building blocks of progress in biomedical research, which include future preclinical experiments and clinical trials. Therefore, the reproducibility of results from these studies is critical for driving scientific progress (1, 2). Scientific reproducibility is rapidly gaining critical importance in biomedical research, which is highlighted by the recent National Institutes of Health (NIH) guidelines called “Rigor and Reproducibility” that aim to improve poor reporting standards and address significant challenges impeding study replicability (2). The lack of appropriate reporting standards and metadata information, such as statistical technique used to estimate sample size and inclusion or exclusion criteria, leads to low reproducibility in biomedical research. A recent analysis of past biomedical research studies estimated that about $28 billion per year is spent on non-reproducible research in the United States (3). Similarly, an analysis of studies carried out as part of the National Institute of Neurological Disorders and Stroke (NINDS)-funded “Facilities of Research Excellence – Spinal Cord Injury” (FORE-SCI) program found that many of these studies cannot be replicated (4). In cancer research, a review of 100 articles identified that only 28% of the papers have correctly reported the experiment techniques, for example randomization of allocation to treatment groups (5). Therefore, there is a clear and urgent need to develop a practical framework that builds on the “Rigor and Reproducibility” initiative to facilitate rigorous study design and scientific reproducibility.

However, there are no existing well-defined frameworks that consist of metadata elements or representation models, which can support reproducible science. Intuitively, provenance metadata, which describes the history or origin of data and incorporates a detailed description of the different procedures used to handle the data, is a natural candidate for supporting scientific reproducibility. Provenance metadata has long been used in computer science to keep track of the origin of data to validate data quality, compute trust values, and support scientific reproducibility. In 2013, the World Wide Web Consortium (W3C), which is the standards organization for Web technologies (e.g., Web Ontology Language (OWL2) and eXtensible Markup Language (XML)), recommended the PROV specifications for modeling provenance metadata. Provenance metadata is often represented using the W7 model describing the Why, When, What, Where, How, Which, and Who of scientific research (6). The W3C PROV specification consists of the PROV Ontology (PROV-O), which uses OWL2 as a formal representation language to model a minimum set of provenance terms (7). As described in the PROV specification, PROV-O is a reference or “upper-level” ontology,
which can be extended to model provenance metadata in various domains, including medical informatics and facilitate the use of provenance information in achieving the objectives of “Rigor and Reproducibility” guidelines.

In this paper, we introduce Provenance for Clinical Research and Healthcare (ProvCaRe) as a flexible framework to support management of provenance metadata and its application in biomedical research. ProvCaRe consists of an ontology model for representing provenance information and software tools to extract, integrate, and analyze provenance information from various sources (e.g., published articles). The ProvCaRe framework is being developed as part of the NIH Big Data to Knowledge (BD2K) program for data provenance. The ProvCaRe framework aims to address three provenance management challenges in medical research: (1) Development of a provenance ontology to model reporting standards for rigorous design of experiments for both hypothesis generation and hypothesis validation; (2) Integration of the provenance ontology into a Web-based provenance software tool to allow search and analysis of previous research studies; and (3) Development of a highly scalable search engine to allow search and analysis of previous research studies for both replication of previous studies and design of new experiments. This paper describes the development and evaluation of the ProvCaRe ontology, which represent provenance information by re-using many ontology classes from existing biomedical ontologies. For example, the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) (8), the Ontology for Biomedical Investigation (OBI) (9), and the Ontology for Clinical Research (OCRe) (10). The ProvCaRe project leverages data and metadata information from the National Sleep Research Resource (NSRR), which is funded by the NIH to create an integrated repository of sleep medicine data consisting of 50,000 studies from 36,000 participants (11).

**Background**

**Reporting Standards for Rigorous Study Design.** In 2014, the NIH organized a meeting involving 30 basic as well as preclinical science journal to develop a set of principles for supporting reproducible science (12). The principles can be broadly classified into three categories: (1) **Statistical Analysis:** describing the statistical analysis techniques used for data analysis to ensure accuracy of results, (2) **Study Reporting Standard:** consisting of 7 topics including terminological standard, number of repetition of experiments under range of conditions, and techniques for randomization, blinding, and sample-size estimation, and (3) **Data Management** as well as sharing (12). These three categories of study metadata form the foundation for the development of ProvCaRe framework. The framework aims to capture essential information to support both rigorous study design and scientific reproducibility. In this paper, we also use these NIH recommended principles to develop the ProvCaRe ontology.

**W3C PROV Ontology.** The PROV Ontology (PROV-O) models common provenance metadata terms and properties using the description logic-based OWL2 (13). The three foundational provenance terms modeled in PROV-O are **Entity**, **Activity**, and **Agent**, which are linked to each other using seven “core” properties, namely `wasDerivedFrom`, `wasAttributedTo`, `wasGeneratedBy`, `wasAssociatedWith`, `actedOnBehalfOf`, `wasInformedBy`, and `used` (Figure 1(a)) (7). Figure 1(b) illustrates the use of these PROV-O terms to create a provenance graph describing the history of an example patient. PROV-O is designed as an upper-level reference ontology that can be used as a common foundation to develop interoperable domain-specific applications using provenance ontologies, such as the ProvCaRe ontology for medical research.

**SNOMED CT.** SNOMED CT is the largest clinical terminology originally created by the College of American Pathologists (CAP) and currently maintained by the International Health Terminology Standards Development Organization (IHSTDO). The 2015AA version of the SNOMED CT with more than 31,000 terms was used in this paper to identify provenance-related classes. SNOMED CT consists of concepts that are connected by relationships, for example **temporal context** to relate occurrence of an event with a time value. The SNOMED CT terms (also called pre-coordinated concepts) can be combined using the SNOMED CT compositional grammar form to represent new concepts (also called post-coordinated concepts). This feature allows SNOMED CT to be used for representing new concepts that are not part of the original SNOMED CT class hierarchy.

![Figure 1: Core terms of the PROV Ontology and example provenance graph](image)
National Sleep Research Resource (NSRR). The National Sleep Research Resource (NSRR) is creating one of the largest repositories of de-identified physiological signals and clinical data elements that were collected from multiple research cohorts and clinical trials (11). NSRR aims to provide free access to about 50,000 sleep studies collected from about 36,000 participants, which include demographic, physiological, clinical, and other types of data. Researchers can use the curated and well-defined sleep data from multiple NIH studies to develop large cohort studies with more diverse population that was not possible with a single study cohort. This is expected to significantly accelerate the pace of research in role of sleep and sleep disorders in pathogenesis of chronic illness such as cardiovascular and metabolic diseases.

Related Work

There has been past work in development of formal models to represent biomedical research informatics, including the Biomedical Research Integrated Domain Group (BRIDG) model (14), the Ontology for Clinical Research (OCRe) (10), and the Eligibility Criteria Extraction and Representation (EliXR) project (15). The BRIDG model focused on developing a representation of clinical trial protocols involving trial design, temporal information, and events. The BRIDG model was used as reference for the development of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). The OCRe project was a comprehensive effort to develop a formal model of study design to support different phases of clinical research, including study design, execution, and interpretation of data. The OCRe model (revision 315 downloaded from the National Center for Biomedical Ontologies) uses formal logic-based OWL2 representation format to represent different categories of research studies, study characteristics, and statistical concepts. The OCRe project developed the Eligibility Rule Grammar and Ontology (ERGO) tool to annotate textual description of eligibility criteria for research studies (16). Similar to the ERGO tool, the EliXR project developed a pipeline to analyze and discover common semantic patterns in eligibility criteria description extracted from ClinicalTrials.gov. The proposed ProvCaRe project builds on these existing projects with addition of data and process provenance concepts to support scientific reproducibility, which is distinct in terms of scope, objectives, and scalability as compared to these projects. In the next section, we describe the use of “Rigor and Reproducibility” guideline template to develop the ProvCaRe framework and its implementation in the provenance ontology.

Method

Clinical researchers often use the Population, Intervention, Comparison, and Outcome (PICO) framework to represent clinical research questions, which are integral to practice of evidence-based medicine (EBM) (17). Many reports have found that formulation of research studies according to the PICO model improves the accuracy and quality of research results and also support better search strategies (18). However, the PICO model lacks support to represent critical metadata information regarding clinical studies to support scientific reproducibility. For example, the PICO framework does not model: (1) the data handling techniques (e.g. inclusion and exclusion criteria for study data), (2) the statistical methods used (e.g. randomization technique), (3) the technique used for data recording (e.g., blood pressure recorded in supine position), and (4) the instruments used for recording (e.g., multi-contact intracranial electrode used for recording electroencephalogram). To develop the ProvCaRe ontology, we followed a three-phase development workflow (Figure 2):

1. **Phase I**: In the first phase, we leveraged the “Reporting standards for rigorous study design”, which is used as the foundation to define the principles of the NIH “Rigor and Reproducibility” guidelines to extend the PICO model with provenance metadata concepts (Figure 2 illustrates the different components of the rigorous study design with PICO model components).

2. **Phase II**: In the second phase, the provenance concepts identified from the reporting standards were normalized and mapped to ontology concepts from existing biomedical ontologies using the National Center for Biomedical Ontologies (NCBO) BioPortal “term lookup” service. The re-use of existing ontology classes reduces redundant ontology development efforts and also facilitates interoperability of the ProvCaRe ontology with existing biomedical ontologies.

3. **Phase III**: In the final phase, the W3C PROV ontology is extended with the biomedical ontology concepts identified in Phase II of the development workflow using the subsumption (rdfs:subClassOf) relationship. The resulting ProvCaRe ontology can be used to annotate medical research study description as part of the ProvCaRe framework to characterize the replicability of results reported by research studies.

**Phase I: Identifying Provenance Metadata Terms from the NIH Reporting Standards for Rigorous Study Design.** Figure 2 illustrates provenance metadata elements representing three categories of information related to medical research studies: (1) Study Design, (2) Study Data, and (3) Study Instrument, which together form the
ProvCaRe S3 framework for supporting scientific reproducibility. The ProvCaRe S3 components include the constituents of the PICO model and the NIH Rigor and Reproducibility guidelines (1). The Study Design component is mapped to the Intervention and Comparison elements of the PICO model, which partially describe the treatment or intervention as well as the comparison approach that is used in a biomedical research study. In addition, the provenance metadata represents the requisite information to describe the technique used to ensure randomization of data to be collected and the allocation of animal or human participants in a study at appropriate level of granularity. Similarly, description of the blinding technique and the statistical method used to estimate the sample size of the study are also represented in the Study Design component. The data handling techniques in terms of inclusion/exclusion criteria used to select data to be used in the research study and the approach used to address missing data elements also forms part of the Study Design section. The details of the method used to design the conduct of the research study, which broadly represents process provenance, are essential for external researchers and groups to replicate results of the study (19).

Figure 2: The three-phase workflow used to develop the ProvCaRe ontology
The Study Data category includes the study data elements used in a research study, which can be mapped to the population and outcome elements of the PICO model. The various study variables used to record different aspects of the research study, for example demography, medication, medical history, anthropometry, and laboratory data describing the population under study as well as the outcome information constitute the Study Data. A systematic representation of the provenance information associated with the study variables, for example the minimum and maximum values for each variable, the threshold used to assign values to the study variables, and valid values, is required to allow results from previous studies to be replicated. The Study Tools category includes provenance metadata describing the details of instruments used to record, measure, analyze, and interpret data in research studies. The instruments used in research studies are a critical component in terms of recording bias due to the sensitivity and accuracy of the instruments, for example the minimum and maximum voltage that can be recorded by intracranial electrodes used to record electrophysiological signal data in neurological disorders. This provenance information can be used to ensure correct interpretation of the study data. Similarly, the characteristic of 12-lead Electrocardiogram (ECG) recordings of systolic and diastolic blood pressure in various sleep studies in the NSRR is important provenance information.

Together, the three components of the ProvCaRe S3 framework can be used to represent the relevant provenance information about medical research studies, which can be used by other research groups to reproduce the original study. Table 1 shows the structural patterns of the S3 framework and the corresponding provenance metadata information derived from the NIH Rigor and Reproducibility principles (SDe, SDa, ST represent Study Design, Data, Tools respectively). The formalization of the S3 provenance terms will allow consistent use and interpretation of the terms in different applications and tools, which can be addressed by incorporating the terms in a provenance ontology. However, instead of creating new ontology terms for the provenance information, it is important to re-use ontology classes from existing biomedical ontologies as recommended by ontology engineering best practices. This will also facilitate interoperability of the provenance ontology with existing biomedical ontologies. Therefore, in the second phase of the development process, we identified relevant classes in existing biomedical ontologies for use in the ProvCaRe ontology.

<table>
<thead>
<tr>
<th>Structural Pattern (S3)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>[SDe] [SDa] [ST]</td>
<td>Provenance describing the randomization method used to assign participants to experiment groups; the demography variables used for cohort identification; the instrument used to record blood pressure of participants.</td>
</tr>
<tr>
<td>[SDe] [SDa]</td>
<td>The protocol used to ensure that investigators are blinded to the intervention before they assess, measure, or quantify experimental outcomes.</td>
</tr>
<tr>
<td>[SDa] [ST]</td>
<td>The use of intracranial multi-contact electrodes to record Stereotactic Electroencephalogram (SEEG) for evaluating patients with epilepsy neurological disorder.</td>
</tr>
<tr>
<td>[SDe] [ST]</td>
<td>The statistical method used to estimate the sample size for a multi-center epilepsy clinical study to identify biomarkers for Sudden Unexpected Death in Epilepsy (SUDEP) using multi contact intracranial electrodes.</td>
</tr>
<tr>
<td>[SDa]</td>
<td>The criteria used to identify outlier values in a research study.</td>
</tr>
<tr>
<td>[SDa]</td>
<td>The demography, medical history, medication, and anthropometry data collected in a research study, their minimum and maximum values.</td>
</tr>
<tr>
<td>[ST]</td>
<td>Provenance describing the features of a multi contact SEEG electrode, for example 10-15 contacts with length of 2mm, diameter of 0.8 mm and the contacts were 1.5 mm apart on the electrode.</td>
</tr>
</tbody>
</table>

**Phase II: Mapping Provenance Terms to Ontology Classes.** Biomedical ontologies represent terminological systems using a formal knowledge representation language such as the OWL2 to enable consistent and accurate interpretation of the terms by both humans and software tools (20). The NCBO BioPortal lists more than 500 biomedical ontologies representing a wide variety of biomedical domains to allow researchers to search and use ontology terms for semantic annotation of biomedical data and also support knowledge-driven analysis of
biomedical data (21). We used the navigation feature of the BioPortal to look up ontology terms representing the provenance terms identified in Phase I of the ProvCaRe development process. We identified the SNOMED CT, OCRe and OBI as the top sources of ontology classes for representing provenance information. SNOMED CT has 19 top-level concepts and many of these class hierarchies contain terms that can be used to the provenance terms described in the S3 framework. For example, the Procedure and its subclasses, including Surgical procedure or Management procedure can be used to represent the different kinds of intervention used in research studies. Similarly, we identified the Social context top-level concept and all its subclasses as appropriate to model the population related provenance information. The Pharmaceutical/ biologic product and its subclasses are appropriate the provenance information of medication and drug prescription of study participants.

In addition to SNOMED CT, the Ontology for Clinical Research (OCRe) (Revision 315) contains many classes that can be used to represent provenance information. OCRe has defined a classification of research studies into two primary categories of Interventional and Observational studies (Interventional study design and Observational study design) (10). The OCRe Statistical concept class and its subclasses, including Statistical method and Variable dimension can be used to represent the data analysis techniques used during analysis of the study data. The Ontology for Biomedical Investigation (OBI) models different components of biomedical experimental processes with a focus on basic science research studies (9). OBI represents both the material and processes involved in scientific experiments, for example specimen and its subclasses, including cell specimen and tissue specimen. The process and its subclasses such as cellular process, disease course can be used to represent provenance of the symptom and diagnostic tests used in studies. In addition, OBI models temporal information, which can be used to represent provenance of various events (e.g., recording timestamps of blood pressure or date of laboratory tests) in research studies. Although these three ontologies included the maximum number of provenance related classes in the BioPortal search results, we plan to use the BioPortal Web services API (21) in future to automatically look up ontology classes for provenance terms as part of the ProvCaRe software tool.

**Phase III: Extending the PROV Ontology to Develop the ProvCaRe ontology.** As we discussed earlier, PROV-O consists of three core concepts of Entity, Activity, and Agent, which are linked together with 44 object properties (including 7 “core” properties), modeled using OWL2 RL (7). We extended PROV-O with provenance terms identified from Phases I and II. The Activity class was extended to model the Intervention, Data Handling, Exclusion/Inclusion criteria, Missing data protocol, and Data collection protocol. The class hierarchy used to model these classes follows the structure and description of the provenance terms described in the “reporting standards for rigorous study design” (1). Similarly, the Population, Comparison, and Outcome components of the PICO model are modeled as subclasses of Study Data, which is in turn modeled as a subclass of PROV-O Entity class. The provenance terms describing the Study Design, which is modeled as a subclass of PROV-O Plan class, are modeled as its subclass, namely Sample size estimation protocol, Blinding protocol, and Randomization protocol. The third component Study tool of the S3 framework is modeled as a subclass of the PROV-O Agent class.

Similar to the class expression approach used to model new concepts in SNOMED CT through post coordination using the SNOMED CT compositional grammar form, we propose to use post coordinated class expression to model new provenance terms using existing ProvCaRe ontology classes. This will allow modeling of a wide range of provenance terms in the ontology without requiring all potential provenance terms to be modeled in the “original” ProvCaRe ontology. For example, provenance metadata describing the method used to record polysomnogram in a sleep research study called Sleep Heart Health Study (S3HHS1), which is part of NSRR, can be represented using the post coordinated expression Polysomnogram | has a | measurement method | of | home overnight recording. To support this post coordination expression functionality, the SNOMED CT relationships will be imported into the ProvCaRe ontology. Figure 3 shows a screenshot of the ProvCaRe ontology class hierarchy.
Results

To evaluate the coverage and mapping functionality of the ProvCaRe ontology, we used study variables representing various aspects of clinical research from 6 sleep studies from the NSRR repository. The data from these sleep studies are being made publicly available on NSRR. A brief description of the sleep studies is provided below:

1. **Sleep Heart Health Study (SHHS1) Study**: SHHS1 is a multi-center cohort study funded by the NIH National Heart Lung and Blood Institute (NHLBI) to determine the consequences of sleep-disordered breathing on risk for coronary heart disease, stroke, all causes of mortality, and hypertension.

2. **Multi-Ethnic Study of Atherosclerosis (MESA) Study**: The MESA study is a multi-center project involving 6 field centers to study the characteristics of subclinical cardiovascular disease (CVD) and risk factors that predict progression to clinical CVD and to predict the progression of subclinical disease in 6815 participants aged between 45-84 years.

3. **Childhood Adenotonsillectomy Trial (CHAT) Study**: The CHAT study is a multi-center, single blind, randomized controlled trial to test whether children aged between 5 to 9.9 years with mild to moderate obstructive sleep apnea randomized to early adenotonsillectomy (eAT) show greater neurocognitive functioning as compared to children randomized to watchful waiting plus supportive care after a 7 month observation period.

4. **Heart Biomarker Evaluation in Apnea Treatment (HeartBEAT) Study**: The HeartBEAT study is a multi-center Phase II randomized controlled trial to evaluate the effects of supplemental nocturnal oxygen or Positive Airway Pressure (PAP) therapy compared to optimal medical preventive therapy over 3 months intervention period. The study involves patients with CVD or CVD risk factors, and moderate to severe obstructive sleep apnea. The primary outcome of the study is 24-hour blood pressure profile.

5. **Osteoporotic Fractures in Men (MrOS) Study**: The MrOS study is a large observational study of the determinants of fracture in older men involving 5994 participants living in communities with age of 65 years or older across 6 clinical centers. In a follow up study, some of the participants were recruited in a sleep study involving unattended polysomnography. The follow up sleep study aims to understand the relationship between sleep disorders, falls, fractures, mortality, and vascular disease.

6. **Cleveland Family Study (CFS)**: CFS is the largest family-based study of sleep apnea worldwide with 2,284 individual enrolled from 361 families and studied on 4 occasions over a period of 16 years.

We selected 124 study variables across the 6 research studies and evaluated them according to two criteria: (1) Coverage of the proposed S3 model in terms of representing provenance of the study variables, and (2) Mapping the ProvCaRe ontology classes to represent the provenance of the variables according to S3. The study variables are not exhaustive with respect to the 6 studies and are only representative of the various categories of data that were considered in the research studies. We classified them into five categories of Demography, Anthropometry, Medical History, Disease Metric, and Medication.

Table 2: The Mapping of the 124 research study variables from the 6 studies using the S3 Framework

<table>
<thead>
<tr>
<th>Category of Variable</th>
<th>Provenance Metadata Components of the S3 Framework</th>
<th>Study Design</th>
<th>Study Data</th>
<th>Study Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demography (Total: 21)</td>
<td>False</td>
<td></td>
<td>True, (Population)</td>
<td>False</td>
</tr>
<tr>
<td>Anthropometry (Total: 12)</td>
<td>True, (Threshold used to select for study)</td>
<td>True, (BMI value, Waist circumference value)</td>
<td>True, (Measurement device)</td>
<td></td>
</tr>
<tr>
<td>Medical History (Total: 74)</td>
<td>True, (End points selected for study events)</td>
<td>True, (Average sleep time, Smoking status)</td>
<td>True, (Diagnosed with Insomnia using PSG reading)</td>
<td></td>
</tr>
<tr>
<td>Disease Related Metric (Total: 15)</td>
<td>True, (Protocol for outlier readings of BP)</td>
<td>True, (BP readings)</td>
<td>True, (12 lead ECG machine used to record the BP)</td>
<td></td>
</tr>
<tr>
<td>Medication (Total: 2)</td>
<td>False</td>
<td>True, (current medications, dosage)</td>
<td>False</td>
<td></td>
</tr>
</tbody>
</table>

Mapping Research Study Variables to the Provenance Components of the S3 Framework. The mapping of the 124 study variables across the five categories to the three provenance components of Study design, data, and tools demonstrates that metadata information regarding many of these variables need to be captured across all the S3 components. For example, the variables categorized into the Medical history, Anthropometry, and Disease related
metrics require the capture of provenance information to describe the design, data, and tools of the research studies (Table 2). The provenance information includes the threshold used to assign participants to the study, for example a research study correlating sleep disorder related breathing with cardiovascular diseases might exclude obese persons with BMI higher than threshold values. Similarly, a sleep research study might include participants who have been diagnosed with insomnia by a clinician to ensure that only participants with clinically validated diagnosis are recruited in the study. In contrast, some of the variables in the Demography and Medication category do not have provenance information corresponding to the Study method component.

The results of this mapping allow us to achieve two objectives: (1) Evaluate the effectiveness of the S3 provenance framework to capture metadata information to support scientific reproducibility; and (2) Create a “provenance checklist” for researchers to ensure that the minimum set of provenance information is collected during reporting of results from a study. This will enable researchers to verify that if the relevant category of the provenance metadata for study variable is missing the study result may not be reproducible. The provenance checklist based on the S3 framework can be implemented in a software tool as a practical application to be used for reporting scientific results that can be reproduced and helps meet the desired goals of the NIH Rigor and Reproducibility guidelines.

Coverage of the ProvCaRe Ontology: Use of Classes to Model Research Study Variables. To evaluate the coverage of the classes currently modeled in the ProvCaRe ontology, we mapped the variables from the 6 research studies to the existing ProvCaRe ontology classes. Table 3 demonstrates that the current version of ProvCaRe ontology includes classes that can represent a majority of the research study variables across the 6 studies used in the evaluation. We analyzed the study variables that were not covered by the existing ProvCaRe ontology classes and identified that a significant number of study variables represented domain-specific information used in a study. For example, the existing ProvCaRe ontology classes do not represent participant responses, for example “frequency of feeling unrested” and “time to fall asleep” in the SHHS1 study. It is important to note that the ProvCaRe ontology is not designed to model domain-specific variables. A more efficient approach to address this challenge is to use post-coordinated concept expressions using pre-coordinated concepts in ProvCaRe ontology to represent these study specific variables.

Table 3: Coverage of the ProvCaRe Ontology to Represent the Study Variables

<table>
<thead>
<tr>
<th>Research Study</th>
<th>Percent Modeled: ProvCaRe Ontology Classes</th>
<th>Percent Not Modeled: ProvCaRe Ontology Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SHHS1 (Total: 72 variables)</td>
<td>15% (e.g., race, ethnicity)</td>
<td>85% (e.g., changes in snoring, surgery as treatment for sleep apnea)</td>
</tr>
<tr>
<td>2. MESA (Total: 14 variables)</td>
<td>50% (e.g., medication, gender)</td>
<td>50% (e.g., trouble falling asleep, average sleep time in main sleep)</td>
</tr>
<tr>
<td>3. CHAT (Total: 10 variables)</td>
<td>80% (e.g., age, ethnicity)</td>
<td>20% (e.g., ever had insomnia, total sleep time)</td>
</tr>
<tr>
<td>4. HeartBEAT (Total: 10 variables)</td>
<td>90% (e.g., systolic and diastolic BP)</td>
<td>10% (e.g., diagnosed with insomnia)</td>
</tr>
<tr>
<td>5. MrOS (Total: 8 variables)</td>
<td>75% (e.g., insomnia severity index, BMI, waist circumference)</td>
<td>25% (e.g., time light off to light on scored as sleep)</td>
</tr>
<tr>
<td>6. CFS (Total: 10 variables)</td>
<td>80% (e.g., gender, BMI, race)</td>
<td>20% (e.g., physician diagnosed insomnia)</td>
</tr>
</tbody>
</table>

It is interesting to note that many of the ontology classes used to represent the 124 study variables in the ProvCaRe ontology are re-used from existing biomedical ontologies, such as SNOMED CT (as described earlier in the Method section). The use of these classes in ProvCaRe ontology demonstrates that re-use of existing ontology classes to model knowledge in a new domain such as provenance is a practical and sound approach and validates the best practices of ontology engineering. In future, we aim to expand the ProvCaRe ontology with additional classes from existing biomedical ontologies through automation of the provenance term lookup service and use of automated
techniques to extract subset of class hierarchy from existing ontologies using open source tools, such as OWLAPI (22). We discuss the proposed approach to use automated ontology extraction and import technique using OWLAPI in the following section together with a description of how the ProvCaRe ontology and the S3 framework can play an important role in supporting scientific reproducibility.

Discussion

Role of provenance metadata in scientific research, including data quality and trusted EHR systems. In addition to scientific reproducibility, provenance metadata has an important role to play in ensuring data quality by allowing users and software tools to validate the quality of data. For example, the increasing automation of data processing and analysis steps using scientific workflows (23) can be leveraged to support automated collection, storage, and subsequent analysis of provenance information. This will allow us to ensure that the quality of data is not compromised as the volume and rate of data generation rapidly increases especially in the context of biomedical Big Data. Similarly, the increasing adoption of Electronic Health Record (EHR) systems across the nation highlights the significant need to track the collection, processing, and modification of patient data as it is transferred across the EHR systems and Health Information Exchange (HIE) platforms. The use of provenance metadata information to keep track of how patient information is modified and used across the healthcare system will allow different stakeholders to ensure that data is secure and it is being used in the intended manner.

The Standards and Interoperability (S&I) Data Provenance initiative by the National Coordinator for Health Information Technology (ONC) is working to incorporate provenance metadata in the Health Level (HL) Common Document Architecture (CDA). This will allow provenance metadata to be used as an integral component of the electronic health informatics infrastructure. Therefore, the development of the ProvCaRe framework with a provenance ontology and related software tools will allow the use of existing provenance standards such as the W3C PROV to be used in provenance-enabled medical informatics platforms and software tools.

Limitations. The current approach to manually select and import classes from existing biomedical ontologies is not a sustainable approach in the ProvCaRe project as the size and scope of the ontology increases with addition of new domains. For example, the ProvCaRe project in currently incorporating neurology data from a large multi-center project in epilepsy called the Center for SUDEP Research (CSR) involving 14 epilepsy centers across the US and UK (funded by the National Institutes of Neurological Disorders and Stroke, NINDS). Therefore, in the next phase of the ProvCaRe ontology development, we propose to use open source OWLAPI to extract targeted segment of ontologies, which we have successfully used in our earlier work to create the Epilepsy and Seizure Ontology (EpSO) (24).

Conclusion

Provenance metadata describing the origin or history of data is critical to support scientific reproducibility. In this paper, we introduced the ProvCaRe framework consisting of three categories of provenance information, namely Study Method, Data, and Tool (S3), and a provenance ontology by extending the W3C PROV Ontology. The ProvCaRe framework is being developed as part of the NIH Big Data to Knowledge (BD2K) data provenance initiative and it aims to formalize the NIH Rigor and Reproducibility guidelines using a 3-phase workflow. The ProvCaRe ontology re-uses classes from existing ontologies to represent critical components of the S3 model for recording provenance of scientific studies. We evaluate the effectiveness of the S3 model and the coverage of the ProvCaRe ontology using 124 study variables from 6 research studies in the NSRR repository. The evaluation results demonstrate that the S3 model captures the required provenance information for the 124 study variables. As part of our ongoing work, we are developing automated approaches to extend the ProvCaRe ontology that will be used for semantic annotation of research studies.

Acknowledgement

This work is supported in part by the National Institutes of Biomedical Imaging and Bioengineering (NIBIB) Big Data to Knowledge (BD2K) grant (1U01EB020955) and the National Institutes of Neurological Disorders and Stroke (NINDS) Center for SUDEP Research grant (1U01NS090407-01).

References

Checklist as a Memory Externalization Tool during a Critical Care Process

Aleksandra Sarcevic, PhD1, Zhan Zhang, MS1, Ivan Marsic, PhD2, Randall S. Burd, MD, PhD3

1College of Computing & Informatics, Drexel University, Philadelphia, PA
2Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ
3Emergency Trauma & Burn Services, Children’s National, Washington, DC

Abstract

We analyzed user interactions with a paper-based checklist in a regional trauma center to inform the design of digital cognitive aids for safety-critical medical teamwork. An initial review of paper checklists from actual trauma resuscitations revealed that trauma team leaders frequently wrote notes on the checklist. To understand this note-taking practice, we performed content analysis of 163 checklists collected over the period of four months. We found nine major categories of information that leaders recorded during resuscitations, including patient values, physical assessment findings, and pre-hospital information. An analysis of types and amount of notes written by leaders of different experience levels showed that more experienced leaders recorded more patient values and physical findings, while less experienced leaders recorded more notes about their activities and task completion status. These findings suggested that a checklist designed for a high-risk, fast-paced medical event has evolved into a dual function tool, serving both as a compliance and memory aid. Based on these findings, we derived requirements for designing digital cognitive aids to support safety-critical medical teamwork.

Introduction

A growing body of work has shown the value of checklists in general hospital care, suggesting that their use can be extended to other clinical domains prone to human error1. Checklists are now commonly used to support a range of complex medical activities, including verifying the steps of safe surgery, performing infection control procedures, and defining daily care tasks in a hospital setting2-4. Trauma resuscitation is a high-risk, time-critical and team-driven process of treating severely injured patients, posing a four-fold higher risk of death from errors than general care5. Due to the fast-paced and complex nature of the process, adoption of decision-support systems and cognitive aids in this environment, including the checklist, has been slow. To attain and retain resuscitation skills, providers rely on individual and team training, but training alone has been found insufficient for ensuring patient safety and long-term compliance with protocols6. Recently, however, several U.S. trauma centers have started implementing checklists in actual resuscitations, showing their positive impact on team performance and protocol compliance7-9.

Although initial reports showed the benefits of checklists in medicine, some follow-up studies have questioned their effectiveness10. Medical work is often characterized as messy, ad hoc, and unpredictable11, yet most medical checklists are static and linear. Checklists typically cover a mix of important and frequent activities, but do not take into account repeated tasks, changing patient and environmental conditions, and less frequent but critical tasks. These checklist limitations have led to both low adoption and poor compliance rates12,13. To achieve optimal performance, particularly in time-pressured and safety-critical settings such as trauma resuscitation, checklists may need to be more flexible and dynamic to allow for adaptability to different patient and user scenarios14.

To inform the design of a dynamic checklist for fast-paced medical settings, we studied the use of a paper-based trauma resuscitation checklist in an urban, pediatric teaching hospital, with a level I trauma center. A checklist is commonly seen as a tool for ensuring process compliance by performing and checking off a set of tasks. We found that trauma team leaders, in addition to checking the items, frequently wrote notes next to those items. The notes contained valuable information for task performance and decision making during trauma resuscitations. This finding suggested that the resuscitation checklist has evolved into a dual-purpose tool to support both the process compliance and leaders’ cognitive functions. In this paper, we performed a mixed-method study to analyze the nature of note taking and other observed checklist use practices, and to derive design requirements for augmenting the checklists to better support complex and highly dynamic teamwork.
The resuscitation checklist was designed to help trauma teams prepare for patient arrival and to guide their performance throughout the patient evaluation. Its goal was to ensure protocol compliance and reduce delays in treatments. The checklist contains 53 items grouped into four sections: pre-arrival plan, primary and secondary surveys, and departure plan (Figure 1). The pre-arrival section lists preparatory steps for patient arrival. The primary and secondary survey sections contain the protocol steps, including the ABCDEs, head-to-toe evaluation, and vital signs checking. The departure plan items facilitate discussion about diagnostic tests and patient disposition. Checklist administration was assigned to the physician leader, a role least likely to be hands-on during patient evaluation.

Figure 1. Trauma resuscitation checklist with physician’s notes.

Methods

Research Site & Data Collection

Our research site was an urban, pediatric teaching hospital and a regional level 1 trauma center. Each year, the center treats about 600 high-acuity and high-risk injured children in the emergency department. Patients are treated in two adjoining rooms dedicated to the initial resuscitation. Each room is equipped with vital signs monitors, large wall displays for augmenting vital signs, and a wall clock and stopclock to help teams manage their time. Additional low-tech decision-support artifacts include a dry erase board for displaying the patient’s weight, trauma flowsheet for documenting the process, wall-mounted charts with patient parameters, and recently—the resuscitation checklist.

Data collection took place over a four-month period in 2012 (05/01/2012-08/31/2012), three months after the checklist was implemented. We obtained checklists from 163 resuscitations that occurred during this time. Checklists were administered by physician leaders, a role responsible for leading an interdisciplinary team of up to 15 specialists, assigning tasks, making decisions, and overseeing the process. This role is usually assigned to an attending surgeon, a surgical fellow, a senior surgical resident or an emergency medicine physician. The leaders rarely participated in hands-on evaluation, and mostly stood at the foot of the bed to maintain overview of the patient, team and resuscitation room. The study was approved by the hospital’s Institutional Review Board (IRB).

Data Analysis

All checklists were first transcribed into a spreadsheet. For each checklist, we transcribed its study ID, the physician leader’s experience level (surgical resident [PGY-4], surgical fellow, surgical attending, emergency medicine physician), weight information recorded in the pre-arrival plan section, list of notes written down by the leader, location on the checklist for each note that was written down, and items on the checklist that were left unchecked. This transcript was then reviewed to resolve any ambiguous and unreadable notes. Two experts in trauma resuscitation also reviewed the transcribed notes for clarity and accuracy.

We first performed quantitative analyses on the checklists to determine the frequency of checked and unchecked items, the number of checklists with and without notes, the amount of notes for each checklist section and item, and the amount of structured and free-form text notes. We then focused on the 125 checklists (77%) that contained handwritten notes. We used an open coding technique to identify and categorize types of information recorded by physician leaders. During the first pass, we identified 50 codes characterizing the notes. These codes were then discussed in a group session to determine which codes to keep, remove, or merge. This process resulted in nine high-level notes categories and 27 subcategories (Table 1). The transcript was updated with these final codes in the second pass to reflect these changes. We also determined the number of notes for each high-level category.
To gain additional insight into note taking, we compared the number and types of notes based on the leader’s experience level by grouping the checklists into those administered by surgical residents and fellows (less experienced) and those administered by attending surgeons and emergency medicine physicians (more experienced).

Findings

We report our findings in three parts. First, we present the results from our general observations of checklist use. We then describe the results of content analysis by categorizing and describing information types that physician leaders recorded on the checklists. We conclude by describing note-taking differences based on leaders’ experience levels.

### Table 1. Notes categories and subcategories, percent of all notes, and examples.

<table>
<thead>
<tr>
<th>Notes Categories and Subcategories</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Values (35%)</strong></td>
<td>Heart Rate (HR) 87 HR; 118; 90s; 19?</td>
</tr>
<tr>
<td>Vitals</td>
<td>Respiratory Rate (RR) 20s; 21; 30; 41</td>
</tr>
<tr>
<td></td>
<td>Oxygen Saturation (SO2) 100% sat; 97% RA; 98</td>
</tr>
<tr>
<td></td>
<td>Blood Pressure (BP) 117/74; 90/palpable; 100 over 65</td>
</tr>
<tr>
<td></td>
<td>Pulse 2+ DP; rad 2+; 2+ bilat.</td>
</tr>
<tr>
<td></td>
<td>Glasgow Coma Score (GCS) 4, 3, 6 13; GCS: 3; 15; 4+4+6=14</td>
</tr>
<tr>
<td></td>
<td>Pupils 5mm sluggish but reactive; 3→2; 4 brisk</td>
</tr>
<tr>
<td></td>
<td>Temperature 36.6 capillary; 37; 37.2 rectal</td>
</tr>
<tr>
<td><strong>Physical Findings &amp; Symptoms (22%)</strong></td>
<td>Location Tender midline; RUQ tenderness; pain L posterior</td>
</tr>
<tr>
<td></td>
<td>Transport Est. Arrival Time (ETA) 10 min</td>
</tr>
<tr>
<td></td>
<td>Demographics Age 2 yo; 7 mo/o; 3 wks; Age 12</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medical History PMH 3x concussions; All: cats; H/O Asthma on steroids</td>
</tr>
<tr>
<td></td>
<td>Previous Treatments (e.g., en route, outside hospital) 100mg fent; CPR; 2 x saline; CT head + neck; PE tubes; intubated</td>
</tr>
<tr>
<td></td>
<td>Mechanism of Injury Fall 5ft; MVC vs Ped; motorbike vs. SUV</td>
</tr>
<tr>
<td></td>
<td>Physical Findings Ø LOC; + LOC; lethargic; alert &amp; awake</td>
</tr>
<tr>
<td></td>
<td>Injuries/Complaints Neck pain; R skull fx; leg bruise</td>
</tr>
<tr>
<td><strong>Care Plan (11%)</strong></td>
<td>Labs Basic trauma lab; tox screen</td>
</tr>
<tr>
<td></td>
<td>Consults Ortho; N Surg consult</td>
</tr>
<tr>
<td></td>
<td>Imaging Tests Chest, c-spine; Xray; CXR; CT Head</td>
</tr>
<tr>
<td></td>
<td>Disposition PICU; Neuro</td>
</tr>
<tr>
<td></td>
<td>Medications/Fluids Bolus fluids; LR 28</td>
</tr>
<tr>
<td></td>
<td>Other Consider NAT; obs pl ambulate</td>
</tr>
<tr>
<td><strong>Injury Type &amp; Location (9%)</strong></td>
<td>L LQ abrasion; multiple lacs; facial bruising</td>
</tr>
<tr>
<td><strong>Task Completion Status (4%)</strong></td>
<td>Deferred; Not done; In progress; N/A</td>
</tr>
<tr>
<td><strong>Treatments &amp; Procedures (3%)</strong></td>
<td>IV/O Size/Location R A/C – 20; 22g L hand; 16 ga</td>
</tr>
<tr>
<td></td>
<td>Type/Rate/Dosage/Time Start LR @ 65; 600 ml LR; NS @ 50; etom in; atropine 15.30</td>
</tr>
<tr>
<td></td>
<td>Intubation Tube Size &amp; Location 4 cuffed, 13 @ lips</td>
</tr>
<tr>
<td></td>
<td>Chest Decompression L needle decompression</td>
</tr>
<tr>
<td><strong>Other (3%)</strong></td>
<td>Notes to trauma office, unreadable</td>
</tr>
<tr>
<td><strong>Laboratory Results (1%)</strong></td>
<td>Dstick = 111; HH values: 50 over 30, 15 over 40, 16 over 48; AST/ALT 117/91</td>
</tr>
</tbody>
</table>

To gain additional insight into note taking, we compared the number and types of notes based on the leader’s experience level by grouping the checklists into those administered by surgical residents and fellows (less experienced) and those administered by attending surgeons and emergency medicine physicians (more experienced).
General Observations about Checklist Use

Our analysis of the frequency of checked and unchecked boxes revealed two groups of checklist items (Figure 2(a)): (1) almost always checked, and (2) rarely checked. The first group consisted of primary and secondary survey items such as “Confirm O2 placement” and “State GCS [Glasgow Coma Score],” which were checked in 80% to 95% of the checklists. Optional items such as “For attending activations” in the pre-arrival plan section and “Prepare patient for travel” in the departure plan section were rarely checked (7%), as was the primary survey item “Give fluid or blood” (14%). These findings suggest that a multi-tiered organization and adaptation of the checklist to specific contexts may be useful. Although prior studies of checklist compliance showed that users sometimes check a box without performing the corresponding task or perform a task without checking the corresponding box\textsuperscript{15}, there is no evidence that physician leaders in our study chose to exclude any specific checklist items. It is therefore appropriate to assume that failures in completing checklist items are uniformly distributed across the checklist, and that the items that were rarely or never checked were also rarely or never performed and vice versa (Figure 2(a)).

We observed several styles of checklist use. Notes were sometimes written next to an item, but the item was left unchecked. In other cases, notes associated with an item appeared in the same checklist section as the item, but not next to it. Physician leaders appeared to prefer free margin space, filling both top and bottom margins with notes on most checklists. We also found a few checklists with all 53 items checked off but without a single note. Items with the largest number of notes were “State GCS” under step D(isability) and vital signs evaluation tasks under Vitals in the primary survey section. Items with the least number of notes included the “If intubating” box in the primary survey section, equipment preparation steps in the pre-arrival section, and departure plan items. A closer look at the relationship between the frequencies of item checking and note taking revealed that for items that were rarely checked, notes were almost never taken (left corner of Figure 2(c)). For items that were frequently checked, the number of notes varied uniformly between 0 and 83 (Figure 2(b) as well as the right side of Figure 2(c)).

We further observed that leaders’ handwritten notes belonged to three groups based on the note structure: (1) free-form text, (2) numeric, and (3) numeric with attributes. Free-form text notes were most common, comprising 63% of all notes taken. These notes appeared in the margins and in the secondary survey section, especially for head and cervical spine exams. The length of free-form notes varied between 1 to 46 characters per note. Most notes (78%), however, were short with fewer than 16 characters. Numeric notes comprised 30% of all notes, mostly appearing next to the vital signs, Glasgow Coma Score (GCS), and temperature assessment tasks. Numeric notes with attributes (e.g., “90/palpable” for blood pressure, “2 mm reactive” for pupil size) comprised 7% of all notes, and were also found next to vital signs, GCS score and pupil size. Here, we observed that the attributes were limited to a small set of choices. For example, leaders wrote one of the following for pupil size: “sluggish,” “brisk,” “reactive,” “blown,” “unreactive,” and “equal”. GCS score notes almost always showed its value broken down into three components: eyes, verbal and motor response. Attributes for the GCS score were therefore implied by the initial of the score component name (e.g., “E3V3M5”) or by the order in which they were recorded (e.g., “4, 3, 6, 13”).

Figure 2. (a) The frequency an item was checked off. (b) The number of times a note was taken for frequently checked items. (c) Scatterplot of the relationship between the frequencies of item checking and their corresponding notes.
Checklist Notes: Information Types and High-Level Categories

Our content analysis of physician leaders’ notes written on the checklists showed that leaders recorded a great deal of information during resuscitations. These checklists (including the leaders’ notes), however, were not part of the official medical record and were usually discarded after the resuscitation, unless kept for research purposes. We identified 27 information types and grouped them into nine high-level categories (Table 1). Below we describe each high-level category in greater detail.

**Patient Values:** Patient values were the most commonly found notes, comprising 35% of all notes written down on the checklists. These values included Glasgow Coma Score (GCS), temperature, pulses, pupil size, and vital signs such as heart rate, respiratory rate, oxygen saturation and blood pressure, and were usually recorded next to their corresponding checklist items (Figure 3). For example, GCS and pupil size values were found in section D of the primary survey, temperature was written in section E, and vitals were written down in the vitals section.

The amount of detail for each value also varied, especially for vital signs. Although most leaders wrote both systolic and diastolic numbers for blood pressure, we also found several instances with the systolic number only (e.g., 100), either alone or with an attribute (e.g., 145-sbp, 90/palpable). Oxygen saturation was sometimes accompanied by “RA” (e.g., 97% RA), meaning that the patient was breathing room air rather than 100% oxygen. Another patient value with varying levels of detail recorded was Glasgow Coma Score (GCS). Most often, physician leaders broke down the score into parameters and then wrote the final score summing up the individual scores, e.g., “4, 4, 6 = 14.” Other, less frequently observed variants of this value included noting only the final score (e.g., “15”) or describing each parameter (e.g., “eye injury, talks, obeys commands”).

In some instances, physician leaders wrote more than one note for the same value. For example, they recorded changes in oxygen saturation levels, as well as more than one blood pressure value (Figure 3). Although no time information was written next to these values, these notes suggested that some leaders used their checklists to record trends in patient data, especially in vital signs. Current vital signs monitors can be setup to show trends but teams do not have time to change the monitor settings and mostly use the default, continuous monitoring of vital signs. Similar findings were observed for the GCS values, with initial GCS assessments crossed out and new values recorded next to them. Although trends in the data were found in only 3% of the checklists, these findings suggest that trends are important during time-critical patient management.

**Physical Findings and Symptoms:** Physical findings and symptoms were the second most frequently observed category, comprising 22% of all notes. These notes were mostly related to findings from the secondary survey and were usually found in the secondary survey section of the checklist (Figure 4). Status for each secondary survey item was written next to its corresponding checkbox. The notes included both normal (e.g., stable, within normal limits [WNL], non distended [ND]) and abnormal (e.g., tenderness to palpation [TTP] diffusely, blood in nares, pain) findings. We also noted a few instances of patient complaints written as part of the findings (e.g., “Pt complains of chest tenderness”). Sometimes, however, leaders would jot down only general status notes using this space, without referring to any of the particular survey components (e.g., “multiple lac to scalp 10.5cm” accompanied by a sketch of the head with lacerations marked). Finally, leaders were observed

**Figure 3.**Instances of recorded vital signs trends: Oxygen saturation (top); Blood pressure (bottom).

**Figure 4.**Notes about physical findings.
specifying location of the findings and symptoms. For example, if an abnormality were found on the patient’s chest, the notes would also specify parts of the chest, as in “RUQ [right upper quadrant] tenderness.”

**Pre-Hospital Information:** Information about incoming patients, such as mechanism of injury, received treatments and patient medical history, serves a critical role in helping medical teams prepare for patient care\(^1\). It was not surprising then to see 12% of total notes dedicated to pre-hospital care, mechanism of injury and patient medical history. These notes were mostly written on the top or left margins, or between the patient label and primary survey section of the checklist (Figure 1). For injured patients coming directly to the trauma center, pre-hospital notes included estimated time of arrival, patient demographics (e.g., age, gender), mechanism of injury, physical findings, and injuries and complaints. For patients that were transferred from another hospital, leaders also added notes about their medical history (e.g., allergies, any chronic illnesses and current medications), as well as the results from previous tests, such as imaging and laboratory tests (Figure 5).

**Care Plan:** Care plan notes included notes on any decisions that leaders made about the subsequent steps in patient care. These notes accounted for 11% of total notes and were typically found in the margins, e.g., the lower right corner of the checklist, below the departure plan section (Figure 6, left). Most of the time, these notes specified laboratory tests (i.e., basic vs. comprehensive labs), imaging tests (e.g., head CT, shoulder x-ray), needed consults (e.g., neurosurgery, orthopedics), and patient disposition. The care plan notes were often presented in the form of an additional list of checklist items, with checkboxes drawn next them (Figure 6, right). This observation suggests that physician leaders who presented their care plan in this form felt that the corresponding items were missing from the checklist and should therefore be considered for a future iteration of the checklist.

**Injury Types and Location:** In addition to physical findings and symptoms, we observed notes about injuries identified during secondary survey. Similar to physical findings, physician leaders noted the location of injuries, e.g., left forehead bruise, R[ight] eye ecchymosis, or abrasion over right eyebrow. Although not as frequent as physical findings, injury notes accounted for about 9% of all notes written on the checklists. As part of these notes, physician leaders recorded injury types such as abrasions, lacerations, or fractures. Most of the time, notes about injury type and location were recorded along with physical findings in the secondary survey section.

**Task Completion Status:** We observed that leaders’ notes sometimes provided a rationale for either checking an item or leaving it unchecked. For example, for “Confirm O\(_2\) placement” in the primary survey section, leaders often wrote “N/A sats 100%” or “100% defer,” implying that oxygen mask placement was unnecessary because oxygen saturation was already at 100%. Or, for preparatory steps in the pre-arrival plan section, one leader crossed the entire section noting that they did not have time to complete the pre-arrival steps because the patient already arrived. Task completion status notes were found throughout the checklist, next to different items, and comprised 4% of all notes.
Treatments and Procedures: Although rare (only 3% of all notes), notes about treatments and procedures represent a significant finding because they suggest that some leaders needed to remember the type and amount of administered medications or fluids (Figure 7, left). Most commonly seen notes in this category were about location and size of intraosseous (IO) or intravenous (IV) placement. These notes typically appeared in the bottom margin of the checklist. Notes about procedures were even rarer because only few patients in our sample required procedures such as intubation or chest tube placement. When found, these notes contained the tube size and location.

Laboratory Results: The least frequently found notes were about laboratory results (3% of all notes). These notes mostly appeared on the checklists for transfer patients because test results came through documents accompanying the patient and were reported upon arrival (Figure 7, right). Some results from basic tests like blood gases and glucose levels were also written for patients that arrived from the injury scene, as these tests could be performed in the trauma bay.

Note Taking Differences between Physician Leaders

To gain more insight into leaders’ note taking, we also examined the differences in types and number of notes based on leaders’ experience levels.

First, we examined the percentage of leaders within each experience level that took notes during resuscitations (Figure 8). More than 70% of team leaders of any experience level took notes on their checklist sheets during resuscitations. We found that 75% of all attending surgeons, 77% of all surgical fellows, 71% of all surgical senior residents, and 83% of all emergency medicine physicians took notes on their checklists. Second, we looked at the average number of notes per checklist by leaders of all experience levels. Attending surgeons were the most prolific note takers, with an average of 15 notes per checklist, followed by emergency medicine physicians (on average, 12 notes per checklist), followed by senior surgical residents (on average, 10 notes per checklist), followed by surgical fellows (on average, 9 notes per checklist). An interesting observation is that more experienced team leaders (attending surgeons and emergency medicine physicians) took more notes than less experienced team leaders (surgical fellows and residents). Finally, we wanted to know if there were any differences in the types of information recorded by physician leaders of different experience levels (Figure 9). We found that more experienced leaders, on average, recorded more patient values (e.g., vital signs, temperature and Glasgow Coma Scores).
and findings from physical assessments (i.e., secondary survey findings), two types of information necessary for
decision making. They also recorded more notes about the patient’s subsequent care plan. On the other hand, less
experienced team leaders recorded more notes about their activities, such as treatments and procedures, and task
completion status. These differences in behaviors may be related to the patient injury severity level. Attending
surgeons act as leaders only during highly acute cases, while fellows and residents lead in less critical cases.

Discussion

Our findings showed that trauma team leaders of all experience levels write down a great deal of information on
their checklist sheets. Initially, this finding was not surprising because prior work has shown that paper records
afford many qualities that are critical in medical work, including mobility, flexibility, and visibility of work17,18.
Notes on paper are also used as transitional records for more efficient communication and coordination19. Upon
closer examination, however, we found that leaders’ notes contained valuable information about patient and task
performance during trauma resuscitations. Although the checklist was primarily designed to assist with protocol
compliance, its use was appropriated for additional purposes during resuscitations, such as memory externalization
and decision making. Prior research on distributed cognition systems has shown that cognitive aids (e.g., maps,
charts, diagrams, paper scraps) have an important role during complex, knowledge-based processes because workers
need to lessen their cognitive load by externalizing information20. These findings suggest that complex medical
processes such as trauma resuscitation require externalizing information from memory, much like navigating a ship
or piloting an airplane20,21. Although these other distributed cognition systems have computerized support for
externalizing critical process information, this kind of support is still paper-based in the resuscitation and other
critical care settings. Paper, however, is static (i.e., checklist items remain the same despite changing patient
conditions) and limited in size, which constrains the number of checklist items; otherwise, it would be inconvenient
to manage a large paper sheet. Also, paper checklists cannot be connected to electronic sources of information,
requiring manual capture of information. Our study offers several implications for computerizing checklists and
other cognitive aids to better support complex medical processes.

Provide Information about Patient Data Values over Time

The most common information type scribbled on the checklist was patient values. Physician leaders needed patient
vital signs, GCS values and temperature externalized to analyze the data and detect any abnormalities. It appeared
that leaders used their checklists for memory externalization, suggesting the need for additional cognitive aids to
support these tasks. Although vital signs monitors can provide information about patient values over time, clinicians
rarely set them to those modes because they preferred the continuous monitoring view. Even so, supplemental
displays in the room could facilitate access to patient data over time by capturing, analyzing and visualizing
information such as trends (upwards, downwards or flat), sudden changes of levels, or spikes in the data.

Make the Checklist Adaptive to Different Patient Contexts and Leaders’ Experience Level

Our findings about the frequency of checked and unchecked items suggested the need for checklist adaption to
different patient contexts. Current resuscitation checklist contains only one complex procedure (endotracheal
intubation), and ignores other major interventions, such as chest tube insertion, chest decompression, or pelvic
stabilization. Because of the limited amount of free space on the checklist and the need to keep the checklist
comprehensive but succinct for easy use, the number of checklist items will always be constrained, regardless of the
medium. The implication here is that only the items that apply to the current patient scenario should be displayed,
which can be achieved through real-time checklist adaptation. As we have observed, some physician leaders
performed “manual adaptation” of the checklist by listing action items under the plan of care as newly added
checkboxes. We also observed a few instances of this approach for items in the treatments and procedures category,
like medications and fluids (Table 1). These observations imply that leaders needed additional checklist items to
indicate the kinds of laboratory or imaging tests they wanted to perform, or to maintain a record of medications and
fluids they administered. An electronic checklist could afford such an extension by adaptive selection of checklist
items or by using multiple tabs. Additionally, items that were rarely checked could be called on demand using the
collapse and expand feature, or shown automatically based on patient-context recognition. Finally, we found that
leaders of different experience levels showed different preferences for types of information, suggesting that some
adaption based on user contexts may also be needed.

Allow Selection-Based rather than Description-Based Note Taking

We found that physician leaders marked not only task completions, but also task outcomes. These observations
suggest that interactions with checklists could be augmented by functionalities that allow easy input of descriptive
information. Writing descriptive notes requires more thought and effort than selecting an item. Documenting during resuscitations, however, requires efficiency due to the rapid information flow, which makes selection-based data entry more appropriate. This type of interaction could be supported by creating and structuring selection lists based on the notes we identified through the analyses. Given that most free-form notes were relatively short (≤15 characters), we expect that the selections lists will have a limited number of items (the shorter the note, the more limited the number of choices of what it can say).

We also observed that leaders sometimes explained the rationale for skipping some checklist tasks. Having a selection-based menu that was structured using their notes could make this practice faster and easier as well. With advancements in sensors and automatic data capture, some of the conditions for skipping a checklist task may be possible to automatically detect and alert the leader that a task is not applicable. For example, as described above, leaders often justified their skipping of the oxygen mask placement by noting that the patient was already at 100% O₂ saturation. Integrating the output of the O₂ saturation sensor into the system could automatically infer that the oxygen mask is not necessary.

Handwritten checklist notes depend on personal styles and preferences. For example, pupil size and response were variously described as “Pupils Equal and Equally Reactive to Light,” “Pupils Equal, Round, Reactive to Light,” and “Pupils Equal, Round & Reactive to Light & Accommodation,” all of which conveyed the same status. By structuring leaders’ observations and providing a selection list, we can ensure standardization across different leaders and checklist use cases.

Allow Free-Form Text Input for Note Taking

The mode of note taking should in general match the note structure. Numeric keypads should be considered for entering numerical notes, such as vitals, GCS, and temperature. Some types of notes cannot be easily reduced to a list of options. As we found, notes about patient medical history and mechanism of injury, as well as questions for the trauma services staff served an important role, even though they were less frequent than notes about task completion and outcome. Future checklists should allow the leaders to write these notes by providing a designated space and free-form text input. Rapid input of free-form text could be further facilitated by selection lists and auto-completion of unique text, which is again possible based on our analysis of note content. Using previously seen notes as completions is feasible because free-form notes are relatively short and it is very likely that the user will find a previously seen match for his or her intended note.

Conclusion

In this paper, we analyzed user interactions with a paper-based checklist in a regional trauma center based on content analysis of physician leaders’ notes scribbled on 163 checklists collected over a four-month period. Our findings showed that leaders recorded nine major categories of information during resuscitations, including patient values, physical assessment findings, pre-hospital information and care plan steps. We also found that more experienced leaders, on average, recorded more patient values and physical findings, while less experienced leaders recorded more notes about their activities and task completion status. These findings suggested that a checklist designed for a high-risk, fast-paced medical event evolved into a dual function tool, serving as both the compliance and memory aid. Based on these findings, we discussed several implications for designing future dynamic cognitive aids for high-risk medical events to better support fast response medical teams.

This paper represents our first look into the checklist note-taking phenomenon. Our future work will continue collecting and analyzing paper checklists to further examine the differences in physician leaders’ use practices based on experience level. We also plan to examine the extent to which factors such as patient scenarios and injury acuity affect types and amounts of notes. Physician leaders will be surveyed to provide their perceptions of the checklist effectiveness, as well as their reasons for note taking. Finally, we will continue refining design recommendations for dynamic checklists and other cognitive aids for complex medical processes.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. #1253285, and partially supported by Health Resources and Service Administration (HRSA) Program Emergency Medical Services for Children (EMSC) Targeted Issues under Grant No. H34-MC-19351. Special thanks to Lauren Waterhouse and Rachel Webman at Children’s National Medical Center for their help with data collection and review.
References

A Novel Conceptual Architecture for Person-Centered Health Records

Titus Schleyer, DMD, PhD1,2, Zachary King, MS3, Zina Ben Miled, PhD3
1Department of Medicine, Indiana University; 2Regenstrief Institute Inc.; 3Electrical and Computer Engineering Department, Purdue School of Engineering and Technology, IUPUI

Abstract

Personal health records available to patients today suffer from multiple limitations, such as information fragmentation, a one-size-fits-all approach and a focus on data gathered over time and by institution rather than health conditions. This makes it difficult for patients to effectively manage their health, for these data to be enriched with relevant information from external sources and for clinicians to support them in that endeavor.

We propose a novel conceptual architecture for person-centered health record information systems that transcends many of these limitations and capitalizes on the emerging trend of socially-driven information systems. Our proposed personal health record system is personalized on demand to the conditions of each individual patient; organized to facilitate the tracking and review of the patient’s conditions; and able to support patient-community interactions, thereby promoting community engagement in scientific studies, facilitating preventive medicine, and accelerating the translation of research findings.

Introduction

Recent, significant investments in health information technology (health IT) have resulted in an unprecedented increase in the “digitization” of the healthcare system in the US. However, the digitization of healthcare information has not automatically translated into increased utility and usability for patients and providers. Indeed, according to a recent report from the Office of the National Coordinator for Health Information Technology (ONC), “there is much work to do to see that every individual and their care providers can get the health information they need in an electronic format when and how they need it to make care convenient and well-coordinated …” 1.

An example of the practical limitations common today is the fact that patients often must interact with more than one patient portal, especially when they see clinicians in multiple health systems. This makes it difficult for them to review and synthesize their health information, and act efficiently and effectively on it. A second limitation is that the health information captured about patients is organized using a one-size-fits-all model. For instance, even in electronic systems male patients are routinely asked whether they are pregnant or not. Thus, in the words of the ONC report, health information is not easy for “individuals, their families, and care providers to send, receive, find, and use in a manner that is appropriate, secure, timely, and reliable”1. Against the backdrop of these limitations, the increasing penetration of health IT, coupled with our - now commonplace - expectation that we can manage almost all aspects of our lives electronically, presents an unprecedented opportunity to shape the health sector via a socially-driven, digital approach for person health record (PHR) management. This paper discusses some of the gaps preventing the mainstream adoption of PHRs and proposes a framework that can help bridge these gaps. The goal of this framework is the automated generation of a PHR customized to individual patients and their health and disease conditions. Through this framework, patients will have the opportunity to manage their own health information based on their specific medical conditions rather than a generic one-size-fits-all model. The expected benefits include:

- consolidating and organizing the underlying data in manner that makes information easily understandable and accessible to patients and health care providers;
- enabling targeted preventive medicine tailored to individual patients in an effective and efficient manner; and
- facilitating large scale studies, and promoting the active engagement of the patient in local, regional and national research.

The Health Information Digitization Roadmap

The roadmap for the digitization of the health sector emphasizes three main initiatives: 1) migrating legacy ad hoc electronic health records (EHR) into standardized enterprise-level EHR systems, 2) facilitating EHR-to-EHR interoperability and exchanges of health information, and 3) enabling EHR-to-Patient/PHR exchanges of this health information.
The success of the first initiative, the migration to standardized EHR systems, has resulted in:

- 93% of non-Federal acute care hospitals adopting certified EHR systems by 2014 according to the ONC2; and
- 200 million patients in the US expected to be covered by the Epic3 EHR system once all of its ongoing system deployments are completed.

With regard to the second initiative, EHR-to-EHR interoperability, the federal government launched an unprecedented nationwide effort in 2009 to help evolve a digital, interoperable health care system in the US4. This effort resulted in more than 50% of office-based professionals and 80% of hospitals using EHRs through which they are now able to electronically exchange standardized patient information in a basic fashion5,6. A similar state-wide EHR interoperability endeavor is the Indiana Network for Patient Care (INPC), which was started by the Regenstrief Institute in 1998 and is now operated by the Indiana Health Information Exchange. Today, the INPC connects 25,000 physicians, 106 hospitals, 110 clinics, surgery centers and many other healthcare organizations7, and maintains information about over 14 million patients.

In connection with the third initiative, EHR-to-Patient/PHR information exchange, it was estimated that as of 2012, 57% of healthcare providers had a patient portal solution in place allowing patients to access all or part of their health information8. Moreover, in order to investigate the challenges of supporting two-way information exchange between patients and health care providers (Figures 1 and 2), the ONC conducted a pilot through the National Association for Trusted Exchange (NATE)9. This pilot focused on data flow and exchange of information from the provider to the PHR (EHR-to-PHR) and vice-versa (PHR-to-EHR). The NATE pilot9 bi-directionally connected EHRs to PHRs using Blue Button10 and HealthVault11. Blue Button allows patients to electronically access their own health information from providers such as health plans, pharmacies, and hospitals. HealthVault is a PHR that allows patients to connect to health providers via Blue Button and download their health related information into a cloud-based account.

The three groundbreaking initiatives mentioned above are essential enablers to support the difficult “last mile” effort that will deliver much anticipated benefits of PHRs to the patient. Currently several patient-oriented software applications12 are being offered as solutions to bridge the last mile. However, the limitations of these solutions are still significant enough to have prompted the ONC to articulate its recent 10-year vision. This vision encourages “last mile” solutions that a) enable individuals to manage their own health information, b) share it with their health care providers and c) enable health care providers to individualize diagnosis and therapy and adapt it, as needed, to the patient’s condition in real-time by 20251.

In the following section, the features as well as the limitations of some of the existing patient-focused solutions are discussed in order to establish a baseline and motivate the need for a PHR framework that is capable to meet the ONC 2025 vision.
Current Patient-focused Software Applications

Different vendors offer patient-focused software applications such as personal health data-marts, process flow and decision support systems. We briefly review these different types of software applications in this section.

HealthVault, Google Health and Indivo-Dossia can be categorized as personal health data-marts. HealthVault provides the ability to organize patient health information (PHI), register medical devices (for acquisition of parameters such as blood pressure and glucose) and aggregate family accounts. Figure 3 shows the specific one-size-fits-all data model used by HealthVault. Under this model, patients have a “health” account on the cloud that is used to house their records retrieved from different health care providers’ EHRs, which are, in the overwhelming number of cases, completely independent of each other.

Google Health was an early PHR that used a model similar to that of HealthVault. It was discontinued in 2011. Indivo is an open source PHR which is supported by several large corporations including WalMart. It allows a limited level of personalization through the selection of apps such as immunization tracker and medication manager. These pioneering tools use a relational model with a focus on collecting data from different sources (Figures 2 and 3).

An additional feature of Indivo-Dossia is the ability to automate the process flow between the health service provider and the patient. For instance, a patient may request an examination by using Indivo. The doctor can then review the patient’s health profile and update it by entering a diagnosis or ordering a laboratory test. The request for the laboratory test is forwarded to the information management system of the corresponding laboratory. Once the test is completed, the results are entered and the doctor is notified. Indivo uses ontologies to semantically disambiguate concepts that may be expressed differently in the participating sub-systems.

MeTree and Health Heritage are decision support systems. MeTree focuses on collecting and organizing family history in order to help support primary care decision-making. Health Heritage focuses on matching the family history with recent scientific research in order to personalize care and prevention plans.

Given the variety and multitude of these types of patient-focused software applications, some of which have been in existence for nearly a decade, the obvious question is: Why is widespread community adoption lacking? The result of a survey of PHR penetration rates in the US indicated that, in 2010, only 7% of the survey participants had ever used a PHR and less than half of these continued using it. Among veterans, 71% of which utilize the Internet, about 20% reported using a PHR in a 2012 survey.

In an attempt to answer the above question and by using a process of elimination, cost can be excluded since most of the above-mentioned patient-focused software applications are free. Privacy concerns can also be excluded based on the findings of an earlier study. This study indicated that while a reasonable level of privacy is expected, a large percentage of the participants were willing to share their health information. The answer can then hypothetically be attributed to 1) the community interest level or 2) ease of use and the added benefit trade-off. In the next section, we present the results of a study that shows that community interest level can also be excluded. This conclusion is supported in the literature with a consensus around the fact that the adoption of PHRs is low in spite of the anticipated benefits. Some of the identified reasons for the low adoption include the need for additional education and training but most importantly for a better PHR model that can leverage technological advances while meeting the needs of patients and health care professionals.

Community Interest

In order to assess the potential interest level, we conducted a simple study to measure the current community engagement for a representative social media platform, Twitter. Twitter activities related to three common diseases, namely cancer, asthma and diabetes, were investigated for three consecutive semesters from 7/1/2014 to 12/31/2015. The results were obtained by querying the tweets for the keywords a) “cancer”, b) “asthma” and “asthmatic”, and c) “diabetes” and “diabetic”. In the case of cancer, the raw data was preprocessed to eliminate irrelevant tweets (e.g., tweets related to the astrological sign cancer).
For each period and disease, Table 1 shows the number of tweets and the number of followers for a given handle (i.e., user) that posted a tweet related to the specific disease during the period. Duplicate entries were removed such that the number of followers for a given handle are counted only once during any given semester. The last column shows the percentage of the total number of followers shared by the top three handles for each disease type.

Table 1 highlights the following general trends:

- The number of followers show a steady increase for all three disease types.
- Cancer shows the highest level of activity, followed by diabetes and asthma. The number of people in the US impacted by cancer (20.3m), diabetes (21.0m) and asthma (17.7m) are relatively close. However, the percentages of total deaths attributed to cancer and diabetes in 2013, are 22.5% and 2.9%, respectively. Asthma is not ranked as a leading cause of death. This comparison could possibly hint at a correlation between the activity level and the mortality rate.
- The top three handles account for 11%, 39% and 30% of the total number of followers for cancer, diabetes and asthma, respectively. This indicates that there are some influencers with large numbers of followers that have the potential to define the landscape of health social media.

The choice of Twitter as a social media platform was only meant as a representative example. Other social media platforms exhibit similarly high health-related activities. Examples of health-focused social platforms include healthboards.com and thepatientforum. Based on the above results, and the fact that Twitter is a widely used social media platform, the interest level can be excluded as a cause for the lack of widespread adoption of PHR systems. Ease of use and the perceived added benefit by the patient seem to be the determining factors in the mainstream adoption of PHR systems.

Conceptual Architecture

As discussed above, one main reason why current PHRs have not yet been able to deliver the benefits articulated in the 2025 ONC vision is likely to be ease of use. Despite the fact that the user-friendly nature of a software application is often associated with the personalization level of the application, most current PHRs have adopted a one-size-fit-all model. The tailor-made, on-demand experience that the community is now expecting from online services (e.g., on-demand TV, personalized marketing and personalized banking) should be reflected in the services provided by the PHR. Furthermore, a key factor in the success of this personalization is the adoption of a condition-driven health data model rather than the widely used event- and institution-based data model. One of the drawbacks of the event-based data model is that, when reviewing, for instance, a patient’s diabetic condition within current PHRs, clinicians must find and navigate to measurements, appointments, medications and procedures among a potentially large set of data points over time. The condition-driven model allows the individualized PHR to be organized around a) the conditions of concern to each individual, b) the potential correlation.
between these conditions, and c) the potential relation between the manifestations of these specific conditions in the
PHRs of the patient’s relatives.

Figure 4 shows the general framework of the proposed PHR framework. The main modules of this framework consist
of a) personalization and registration, b) data management, c) services and updates, and d) notification and
authorization modules. Each of these modules is discussed next.

a) Personalization and registration

The concept of personalizing the PHR by using a condition-based model is sustained throughout the life-cycle of the
PHR. When patients initially register, the personalization and registration module will dynamically and interactively
create a set of questions enabling the rapid capture of the patient’s health status. For instance, knowledge of the gender
of the patient will eliminate, in either case, a large number of gender-irrelevant questions. Similarly, entry of the date
of birth will allow the registration to be tailored to health conditions pertinent to the appropriate age group. In addition
to its efficiency, this approach offers the possibility to drill down and gather information for important health
conditions comprising historical information related to a given disease, family members suffering from the same
disease, and previous and current health care providers related to this condition.

The interactive and dynamic registration questionnaire is only the first step of the personalization process. The next
step consists of generating a personalized PHR that will organize the health information provided both by the patient
as well as that obtained through the healthcare provider in a semantically meaningful way. The details of this
personalization are covered throughout the remainder of the paper.

b) Data management

Our proposed data management scheme is based on MongoDB28, a NoSQL document database. MongoDB stores data
as documents by using an extension of JavaScript Object Notation (JSON) called Binary JavaScript Object Notation
(BSON) which includes primitive data types such as floating point. Most of the current EHR-EHR data exchanges
today transmit data directly through HL7 or use the Extensible Markup Language (XML). The debates about the
suitability of XML versus JSON or its variant BSON is ongoing. Our choice of BSON for both data representation
and data exchange is based on the fact that the latter supports arrays which is an essential data structure for the
representation of repetitive measurements especially when the data has a high velocity (e.g., reading of glucose levels
multiple times a day). Furthermore, the conversion from XML to BSON is straightforward and can be easily
accommodated to retrieve PHI from existing EHRs.

The use of the label “document” in MongoDB is a misnomer. In this paper, the term object is used instead. The
proposed data model is designed around basic primitive objects (i.e., documents in MongoDB technical terms). These
primitive objects include basic information, measurement, procedure, lab test, prescription and relation. The primitive
objects are subject to update and extension without incurring any disruption to the operation of the application. This
is a unique feature of MongoDB which simplifies the personalization process. For instance, it is conceivable in the
future to add a meal object that will support the nutrition aspect of the PHR as well as the individualization of the
managed delivery of insulin for a diabetic. Another possible extension is a physical exercise object that will be used
to support fitness activities.

Figure 5 shows a sample instantiation of four of the primitive objects, namely basic information, measurement,
observation and relation. These primitive objects are patient-specific and generated by the registration and
personalization module (Figure 4). All of the objects reference a “uid”. This is a unique health id for the patient. In a
previous study19, participants indicated that they did not object to the use of a unique health id. Furthermore, most of
the current users of social media platforms are accustomed to having a unique identifier. For instance, in Twitter this
identifier is called a handle and it was used to generate the statistics of Table 1. The health care provider is also
associated with a unique identifier (Figure 5-2), with the existing National Provider Identifier being the most logical
choice. Moreover, all measurements, observations and scanned documents also have a unique identifier. These latter
identifiers are constructed by prepending the uid of the patient to a unique sequence for each element type.

Several of the features of the primitive objects are worth noting:

− The measurement type includes a potential list of mappings. These mappings are added as needed to each individual
patient record in order to support interactions with health care providers that use different vocabularies. Mapping
between different vocabularies can be performed by using a metathesaurus such as the UMLS29. Furthermore, on a
longer term basis, this field can also be used to capture any vulgarization of different measurement names. In general, this functionality reduces the difficulties associated with semantic mapping.

- The primitive measurement object includes an aggregation type field which indicates how the data should be aggregated (e.g., sampling, monthly averages, cumulative averages, etc.). The aggregation type allows the definition of the most appropriate summative method for high velocity data. The method is customizable to each individual, condition and measurement. For instance, different conditions in the same patient may rely on different aggregations of weight or blood pressure readings.

![Sample Primitive Objects](image)

**Figure 5.** Sample Primitive Objects

- The reading sequence is a sequence of tuples consisting of a timestamp and a value. The sequence is updated with every reading associated with a measurement. Furthermore, a new measurement object will be instantiated if any of the underlying fields such as the ordering party or the device id are changed.

- The observation object is used to capture the results of an encounter between a patient and a health care provider. Depending on the result of the observation, the PHR may be restructured to highlight a new condition or the observation may be linked to an already existing health condition.

- The relation object will help identify the type of relation (e.g., parent, sibling, child, etc.) as well as the associated uid of the relative. The underlying application will then perform regular updates based on routine updates to the relative’s PHR or based on newly discovered research to establish the association among any of the individual’s
health conditions and those of a relative. Furthermore, a scoring mechanism is established and dynamically refined to indicate the strength of this association with each of the relatives and for each of the specific health conditions. Several indexing mechanisms are available in MongoDB that support this type of association.

c) Services and updates

Each individual interacts with the server and the health care provider through her personal digital assistant (PDA). Both the server application and the PDA-application have a three-tier architecture: 1) the front-end is implemented using HTML, CSS and JavaScript, 2) the middle layer is based on GoLang, and 3) the back-end uses MongoDB. On the server, MongoDB is used to store all the data related to the registration, personalization, authentication, services and updates. It also manages user profiles and displays data according to each user profile.

There are two main types of profiles: the patient profile and the health provider profile. The health provider profile is outside the scope of this paper. It essentially classifies the health providers into roles including physician, nurse, insurance, institution, etc. The physician profile allows the user to interact with the patient through messaging and enables the physician to organize the patients into different groups (e.g., a “critical watch list.”) The physician, once authorized by the patient, can also place the patient’s measurements under alert. This configurable functionality will enable the physician to receive an alert, for example, when the glucose level of the patient exceeds a certain level. The interaction between the health provider and the individual described above represents one layer of communication. The other communication layer consists of the interaction of the individual with the community.

One of the functionalities of the services and updates module in Figure 4 is to extract health information from different sources of specific interest to the patient. This information can be the result of recent research, a nutrition database or preventive health information. This functionality is key as it will help the timely translation of scientific findings into immediate benefit to the patient thereby promoting retention and mainstream engagement.

The services and updates module will also be responsible for recruiting participants for large scale studies. In this case, a given uid will send a request to the server for the distribution of a recruitment form to participants. The server will review the request and based on the uid’s trust level and a pre-authorized protocol will either decide to forward the request to the target participants or request additional endorsement/validation. Once this authorization step is completed, the server application will generate a list of qualified participants. Additional restrictions may apply during this step. For instance, the requesting party may be asked to refine the underlying query if the number of participants that qualify exceeds or is below a given threshold leading to resource exhaustion or privacy violation, respectively. This automated data validation process is currently under investigation.

Once the validation process is completed, the services module will contact the participants using a push notification technology and will register the uid of the users that accept to participate in the study. The anonymized and sanitized data corresponding to the participants will be extracted and forwarded to the requesting investigator. The investigator will be required to disclose the result of the study which will be made available to the participants and to the community at large. Furthermore, the users will be able to score these findings in terms of benefit to their health, thus impacting the accumulated contribution score of the individual investigator. This type of engagement allows the patients to feel empowered and is critical to their retention.

Most of the activities related to the engagement of the patient in community studies and research are handled by the server application without any involvement from the patient beyond the authorization and the feedback steps. In order to further facilitate this process, the user may decide to setup general rules that are applicable to all requests in a given category. An example rule may be to accept requests for participation for all studies sponsored by a given institution or a given individual. This type of rules promotes “brand” loyalty and facilitates stronger research-community relationships.

Typically, HIPAA prevents direct researcher-patient contact and mediation through a health care provider is mandated unless the participant clearly opted in to direct contact. In the proposed architecture, the patient has to provide prior authorization. Beyond this step, the exchange of information is mediated by the services and updates module which will have to be under the administrative management of a “trusted party”. The policy implication of this data exchange model and other models proposed in the paper is the subject of future investigation.

d) The notification and access authorization module

As previously mentioned the structure of the communication of the proposed framework consists of the combination of two layers. The first layer is a cloud server and supports the communication between the patient and the community.
This layer has a centralized architecture and includes the notification and access authorization module shown in Figure 4 which is responsible for:

- The authentication of all users, including patients, service providers and researchers.
- The management of patient-health care provider access authorizations which are given on a per condition basis by the patient to the health care provider. Connection rules are managed through the front-end and data access rules are managed through the back-end in MongoDB.
- The management of access authorizations required for the participation of the patient in research studies.

The second communication layer is a Peer-to-Peer network as illustrated in Figure 6. Peer-to-Peer networks are becoming increasingly popular in social computing (e.g., Napster). In this layer, peers are autonomous and self-organize in a private sub-network that co-exists with other patient-centric or provider-centric networks. Once a user is authenticated by the notifications and access authorization module in the first layer, she can exchange data directly with her selected peers without the involvement of the server. This mode of communication is used for the exchange of private data between the patient and the health care provider. In addition to scalability and congestion control, the Peer-to-Peer network offers better data exchange security.

The Personalized PHR

The role of the registration and personalization module is to generate the specific meta-model for the PHR and to populate it with the appropriate data by using the other modules including the services and updates module mentioned above. Figure 7 shows a high level representation of this model for a specific individual. The skeleton of the model is generated by the registration and personalization module when the patient first interacts with the platform. This model is then enhanced and updated as it is enriched by the user, the health care provider and the users’ network.

The services and updates module will rely on available application program interfaces (APIs) in order to extract the information needed from different EHRs and research data sites. The APIs can be triggered on a regular basis to update the PHR record of each individual patient. These data are stored in the MongoDB database by using the primitive objects and consolidated based on the uid. A copy of these data and the meta-model are available as part of the back-end database of the client application on the patient’s PDA.

In most traditional client-server applications, the client-side application retrieves data on-demand from the server. In the proposed architecture, the client and the server are both responsible for different types of data exchanges and services. The server application is responsible for the exchange of information between the patient and the community. The client application is responsible for the management of the patient’s private health data. For instance, direct capture of measurement from a home health device and its delivery to the health provider will be handled by the PDA through the patient-centric Peer-to-Peer network; whereas, the delivery of health data...
necessary for a study will be handled by the server application once patient approval is granted. This delineation allows the patient to focus on the health management issues that are of most concern to his condition. Furthermore, it facilitates managed control of time-critical exchanges with the patient and his personalized private health network (Figure 6).

**Data Retrieval and Indexing**

One of the major challenges of the proposed application is the ability to provide ease of access for a specific individual as well as across a large set of individuals. Most current database applications either provide one or the other. For instance, organizations often segregate among the two types of accesses by using a transactional database (e.g., EHR), as well as a data warehouse for analysis and reporting purposes. Combining the two is the subject of ongoing research. One of the emergent application of this combination is social computing. The social media model is relatively simple and based on relations such as “is a friend”, “is a follower”, etc. The proposed PHR model also needs to support both transaction-based interactions (e.g., health care provider – patient) as well as analytics-based processing for large scale studies. However, the health context makes the data retrieval and management complex. Therefore, communication has been segregated into two layers, services have been carefully assigned to the client or the server and data management has been associated with the most appropriate application depending on the context.

Patient local data are essentially managed by the client application and then bi-directionally synchronized with the server as needed. For high velocity data, aggregation can help reduce the size of the data that needs to be stored on the PDA. A detailed view of the data can be obtained on-demand through the server application. Although MongoDB is not an SQL database, it does support aggregation which can be used to implement the above functionality. The other type of data that may represent a burden for the client application is the document type (e.g., images). In this case, the original copy of the document can be stored on the server and the client application can access it on demand from the server using a hyperlink.

Because of the potentially large volume of data on the server, proper indexing becomes a critical requirement. Data on the server is organized by using collections of primitive objects (e.g., basic information, measurements, observations, relation, etc.). These collections are for all individuals and indexed by the uid. In order to facilitate large-scale queries, several other indexes (e.g., indexes by condition or geolocation) need to be maintained. Furthermore, compound indexes that are composed of more than one predicate will be developed overtime as the system matures and is enriched by complex user queries. These indexes may eventually hold the most interesting answer to large scale research questions as they will be able to associate, for example, individuals with different phenotypes or genotypes.

Tree and geospatial indexes can also be used to organize the relation of an individual’s PHR to that of her relatives’ PHRs, thus facilitating condition-based distance scoring. Finally, the Time-to-Live (TTL) index will be used to adequately manage the data in MongoDB. The TTL index is customarily used to automatically delete old data from the system. In this case, the TTL index will be used to archive data to a secondary database. This multi-tier database stack allows a) patient centric data to be close to the patient on his PDA, b) community current data to be close to the researchers and c) historical data to be available for long term studies.

**Conclusion**

The proposed platform will help close important gaps in the healthcare research, clinical practice and patient continuum by allowing patients to better manage their health, and benefit from up-to-date research efficiently and effectively. The proposed framework will link a patient’s health information to relevant health data about their family members, automatically determine the applicability of current medical evidence or recommendations to the patient, and enable the patient to enrich his record from external data sources. We believe that these features will facilitate mainstream adoption and, more importantly, sustained engagement by the patient in managing his health.

The innovative aspects of the proposed framework and its departure from the limitations of current approaches center on:

- developing methods to custom-build on-demand applications that will manage data for each individual patient in a personalized and focused manner; and
- identifying mechanisms that can translate research findings into practical and systematic methods for organizing patient health information.

We believe that this level of quality of service will keep patients engaged and interested in the management of their data. Furthermore, it will encourage patients to become willing participants in large-scale medical studies that can
further improve research. Our goal is to leverage the viral effect of social networks towards a customizable solution for patient-centered health record and a cost-effective mechanism to provide answers to important questions that are unattainable through traditional research studies.

References

1. ONC. A 10-Year Vision to Achieve an Interoperable health IT Infrastructure. Office of the National Coordinator for Health Information Technology 2015.
5. ONC. Data Analytics Update. Office of the National Coordinator for Health Information Technology 2014.
10. ONC. About Blue Button. Office of the National Coordinator for Health Information Technology 2015
23. CDC, National Data, Center for Disease Control, 2013.
25. HealthBoards. Health Message Board. healthboards.com
30. Splice Machine. First RDBMS Powered by Hadoop and Spark. splicemachine.com
Adoption of a Nationwide Shared Medical Record in France: Lessons Learnt after 5 Years of Deployment

Brigitte Séroussi, MD, PhD1,2,3, Jacques Bouaud, PhD1,4
1 Sorbonne Universités, UPMC Univ Paris 06, INSERM, Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, Paris, France
2 AP-HP, Hôpital Tenon, Département de santé publique, Paris, France
3 APREC, Paris, France
4 AP-HP, DRCD, Paris, France

Abstract

Information sharing among health practitioners, either for coordinated or unscheduled care, is necessary to guarantee care quality and patient safety. In most countries, nationwide programs have provided tools to support information sharing, from centralized care records to health information exchange between electronic health records (EHRs). The French personal medical record (DMP) is a centralized patient-controlled record, created according to the opt-in consent model. It contains the documents health practitioners voluntarily push into the DMP from their EHRs. Five years after the launching of the program in December 2010, there were nearly 570,000 DMPs covering only 1.5% of the target population in December 2015. Reasons for this poor level of adoption are discussed in the perspective of other countries’ initiatives. The new French governmental strategy for the DMP deployment in 2016 is outlined, with the implementation of measures similar to the US Meaningful Use.

Introduction

In 2000, the Institute of Medicine published a report about medical errors “To Err is Human: Building a Safer Health System” estimating that up to 98,000 deaths in the US may occur annually as a result of medical errors. This alarming number of deaths was debated, some people suggesting that it was over-estimated. In 2013, James updated the estimates about deaths caused by medical errors and reported that “the number of premature patient deaths associated with preventable harm could be as many as 400,000 per year”. No country is immune to the increasing number of avoidable premature patient deaths. Indeed, patient management has become critically complex. Patients have more and more often various pathologies and comorbidities that require input from multiple healthcare providers across many different care settings. If care is not adequately coordinated among the different providers in charge of a given patient, the quality of care experience for her and for her family can be frustrating at best, harmful at worst. Numerous studies have reported that quality improvement strategies focused on the coordination of care reduced hospital admissions among patients with chronic conditions other than mental illness and reduced emergency department visits among older patients.

The US AHRQ defines care coordination as “the deliberate organization of patient care activities between two or more participants (including the patient) involved in a patient’s care to facilitate the appropriate delivery of health care services”. However, care coordination goes beyond organized care activities. For instance, unscheduled patient encounters across providers escape the organized care coordination framework. Non-deliberately organized care activities may indeed occur when patients move or travel, or in case of accidents or emergency situations. The management of patients in such situations requires that information about their medical condition (current and past) be available to any “new” healthcare provider to ensure care quality and patient safety, and guarantee the continuity of care, e.g. medication reconciliation. This is of primary importance in emergency situations where the lack of knowledge on a given patient condition could yield unexpected adverse events, especially when the patient herself cannot provide vital information or current medications during the medical interview (unconsciousness, mental disabilities).

A solution to improve medical information sharing is digitalization that makes health information “‘liquid’ enough” to flow across healthcare institutions and providers. In most countries, policy makers have promoted the adoption
and use of electronic health records (EHRs) to address health information sharing and care coordination. Medical information is thus available to healthcare providers of the same care team, which is usually the case of healthcare providers within the same institution, from a single practice to a large hospital, or to wider managed care consortia. Clinical information remains locally stored in original health information systems, but may be accessed on demand through data exchange protocols. This is the model adopted by the US which promote health information exchange (HIE) between different EHRs.\textsuperscript{8} The challenge of this model is to achieve interoperability between multiple EHRs. The US Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 has offered substantial financial incentives for providers demonstrating meaningful use (MU) of EHRs. In this framework, virtually any patient data available in EHRs can be collected, whether it comes from affiliated physicians’ offices, hospitals, and clinics, or from completely disparate systems. In this way, clinical information follows patients as they move across different care settings, whether or not they share an organizational affiliation.

However, due to the challenges of standardization and interoperability, networking local EHRs through HIE to get a holistic view of a patient condition remains uncertain. An other solution to computerized medical information sharing is represented by a centralized framework where care records, centrally-stored in purpose-built platforms, are specifically created to serve information sharing. They usually do not contain the comprehensive information stored in all EHRs but only what is considered as necessary to support a coordinated management of the patient inside and outside care coordination practices. This information may be automatically extracted from EHRs or manually added. When a healthcare provider needs some medical information for a given patient, he may access the centralized care record with the consent of the patient. One of the most advanced example of such a centralized nationwide system is the electronic summary care record (SCR) implemented in the UK.\textsuperscript{9} In Scotland, the SCR contains elements such a patient’s name, address, age, allergies, current medications, diagnoses. It is connected to 100% of general practices, and is automatically updated from existing GP records.\textsuperscript{10} Its main objective is to be accessed in emergency and unscheduled care scenarios.\textsuperscript{11}

Based on the UK experience, Eason and Waterson\textsuperscript{12} reported that the centralized model is less feasible than the distributed model. In addition, it relies on a classic “top-down” approach to e-health whereas a “middle-out” approach is more likely to succeed.\textsuperscript{7} Quantin et al.\textsuperscript{13} also advocated for a distributed model of connected EHRs on the basis of pragmatic and cost reasons. On the contrary, Lapsia et al.\textsuperscript{14} suggested that the centralized model for nationwide health information infrastructure scales better than the distributed model in terms of data availability and integrity, with lower failure rates. However, literature about centralized nationwide care records generally reports poor actual adoption.\textsuperscript{11,15} The same difficulties in the adoption and deployment of HIE are reported with the distributed model despite MU incentives.\textsuperscript{16,18} A recent study reported that even among US HIT/EHR adopters, one third does not routinely receive the needed patient information for care coordination.\textsuperscript{17} Thus, sharing patient information from distributed locations through HIE has been questioned, and initiatives aiming at “recentralizing” the model have been launched. For instance, health record banking promotes the centralized storage of and access to care records fed by HIE. Such health record banks are not operated by governmental bodies and might rely on private organizations.\textsuperscript{18}

Beyond digitalizing medical information, another solution to improve the sharing of the same holistic view of a patient’s condition is to count on patients and their caregivers to serve as care coordinators, filling information gaps between providers, addressing conflicting instructions, and managing redundant orders and medications during care transitions. This is possible only if patients are informed, involved, and engaged. The Blue Button project has become a way for many Americans to download their health records. A recent survey reported that one third of patients at the VA were Blue Button users.\textsuperscript{19} The most highly endorsed benefit was that it helped patients understand their health history and thus be active in their management.

In 2004, the French government decided to develop a “personal” nationwide care record, called DMP or “Dossier Médical Personnel”. First elaborated as a set of distributed region-wide records, the DMP was re-engineered in 2009 using the centralized model and was officially launched in December 2010. Though personally-controlled to promote patient empowerment and built according to the centralized model to avoid interoperability-related technical issues, the DMP has had a chaotic history. This article provides a description of the historical background of the DMP in France, the rates of adoption of the DMP nationwide, an analysis of the reasons why, five years after its launching, DMP adoption rates did not match expectations, and the new solutions to be implemented to re-launch the program.
Background

Ups and downs of the DMP. The French health system is a centralized public system controlled by the government (Ministry of Health) with a regional support (Regional Health Agencies in charge of the regional implementation of National health policies). The government health insurance (in French “Assurance Maladie”) funds most care expenses of the population, mainly from the social contributions of active workers. All workers (either currently active or retired) benefit from the government health insurance: they are namely insured. Their health care coverage includes both their own medical expenses and the care costs of their spouse, when unemployed, and children. Care is provided in medical institutions, either public or private, monitored by the French Ministry of Health. Outside health care facilities, private self-employed health practitioners are monitored by the government health insurance (patients used to pay their doctors for medical consultations and were reimbursed by the government health insurance, but now the government health insurance directly pays the medical consultations to doctors). Some National pay-for-performance programs such as ROSP (“Rémunération sur Objectifs de Santé Publique”) have been implemented by the government health insurance that offer financial incentives to private physicians for meeting certain performance measures (29 criteria). Adoption of an EHR is one of these measures and yields €525 (US $580) per year. When fulfilling all criteria, private doctors may earn an additional bonus up to €6,000 (US $6,650) per year.

The DMP has been created in France by the Act of August 13th, 2004 reforming the government health insurance. Only individuals namely insured by the government health insurance may have a DMP. The DMP was expected to improve the quality of care, and to reduce health costs by €3.5 billion (US $3.9 billion) yearly. In 2005, a public interest group, the GIP-DMP, was set up under the supervision of the Ministry of Health, in order to manage the design and the implementation of the DMP throughout the country. The first DMP was relying on a distributed model involving regional medical records. From 2006, several experiments were launched in different regions. These experiments permitted the mobilization of the different stakeholders (healthcare practitioners, patients, and medical software vendors) on the strengths and weaknesses of digitalized data, and the need for legal and technical safeguards to preserve privacy. In 2007, serious concerns about the safety and the confidentiality of personal medical data recorded in DMPs were raised and the program considered as potentially unethical was stopped.

After numerous audits to determine how the program could continue with appropriate security warranties, the DMP was re-designed in 2009 as a nationwide centralized personal medical record with a focus on the protection of patients rights. DMP regulations were written in the Law (Code of Public Health) as stated by the HPST Act (in French, “Hôpital, Patient, Santé et Territoires”) confirming the DMP as a National public health program to serve patient empowerment. A national agency for the promotion of shared health information systems (ASIP Santé) was created to develop and implement the tool. At that time, it was forecasted to get 2 millions of DMPs the launching year (2010), 5 millions the second year (2011), 9 millions the third year (2012), 11 millions the fourth year (2013), and finally 13 millions at the end of the first cycle (2014), thus covering one third of the ultimate target population (38 million of individuals namely insured by the government health insurance, out of the 66.6 million inhabitants of France). However, the adoption of DMP was very slow. In 2012, a report from The Court of Auditors (in French “Cour des Comptes”), an independant public body in charge of conducting financial and legislative audits of most public institutions, proclaimed that the DMP costed about €210 millions, for only 150,000 records (in July 2012), and severely criticized the management of the program. This report almost killed the DMP program.

Model and Functionalities of the DMP. The DMP is a centralized, nationally shared, widely accessible, patient-centered electronic medical record. It is optional and free for all the individual recipients of the government health insurance. DMPs are created by health practitioners only after the patient gave her informed consent (opt-in model). The creation of a DMP requires that both the health practitioner and the patient are “strongly” authenticated. Currently, this strong authentication is achieved by the electronic reading of the smart identity card of the health practitioner and of the health insurance card of the patient.

Patients may access their DMP via a web-accessible DMP portal (patient access). They have to identify themselves using the login-password they were given at the moment their DMP was created.

Health practitioners involved in the management of a given patient need to be authorized by the patient to read infor-
ation recorded in the DMP. For accessing, reading, or pushing information in the DMP, health practitioners need to be strongly authenticated (health practitioner identity smart card or equivalent). However, unauthorized health practitioners may “break the glass” and then access the DMP without the patient’s consent only in case of emergency, when the patient is unable to give her consent. Similarly, practitioners must be strongly authenticated to break the glass.

Health practitioners may access their patients’ DMP through their own EHR. When the EHR software is DMP-compatible, both records are interoperable and smoothly interconnected. DMP information appears in the EHR interface and documents are tagged according to their origin. Figure 1 shows a screenshot of an EHR with the display of the list of the patients of the day in the left window, with a green DMP logo when the health practitioner is authorized to access their DMP, a grey DMP logo when she is not authorized, and no logo when patients don’t have a DMP. In the upper right window, news concerning some patients (that may be either planned for encounter on the same day or not) are displayed: new documents may be available in the DMP (green DMP logo) or have been received by secure messaging (grey envelope). Abnormal results are in red. DMP-compatibility relies on the interoperability framework defined by ASIP Santé and based on international standards (HL7 CDA r2).

![Figure 1. Screenshot of the home page of a DMP-compatible EHR for health practitioners (courtesy of ASIP Santé).](image)

When the EHR is not DMP-compatible, health practitioners may use the DMP portal (professional access) to access, read, and post information into the DMP of their patients. Once again, they need to be strongly authenticated. However, in this case, they have to manage two different environments, their EHR on the one side, and the DMP on the other side, which is known to be less convenient.

As opposed to the UK where patient information is automatically updated from GP records in the Emergency Care Summary, updating the DMP content is an active process: health practitioners are required to push into the DMP the documents they consider to be relevant for care continuity. With a DMP-compatible EHR, the list of the documents newly created for a given patient is displayed on the screen for the health practitioner to check if they should be shared in the DMP. She has then only to tick a box to select the documents to be sent to the DMP.

Because all the operations on a DMP require the operator to be strongly authenticated, the system logs each access to the DMP (name, date, time, break glass procedure, consultation, addition of a document, etc.) and this information is available to the patient as an audit log. The DMP owner and her primary care physician may see all DMP accesses.
Thus, the patient can check that only authorized health practitioners accessed her DMP. Other health practitioners than
her primary care physician only see their own traces. Unauthorized access is subject to criminal prosecution.

If the DMP is not expected to be reduced to a single patient summary as it is the case in UK with the SCR, it does not
aim at being exhaustive and providing the complete picture of a given patient medical history. The DMP should contain
only the relevant information a health practitioner considers necessary for other health practitioners in order to provide
efficient and secure care, be it scheduled or not. A non-exhaustive list of documents that may be posted into DMPs
according to best practices is provided. This list covers the main medical reports: updated patient care summary
(elaborated by the primary care physician), current prescriptions of drugs and care, hospital discharge summaries,
imaging reports, laboratory tests.

Currently, the DMP is essentially document-based with no structured medical data except some metadata (e.g. doc-
ument type, date of production, author). Documents may have three different statuses. They may be either open,
hidden, or sensitive. When they are open, they are accessible to all authorized health practitioners. When they are
hidden, they are only accessible to the author of the document and to the primary care physician. In this later case,
documents are “invisible” for all other authorized practitioners: they don’t even know that some documents they don’t
have access to exist (the hiding of the document is masked). Sensitive documents are documents that are temporarily
not accessible to the patient. This concerns for instance the anatomo and cytopathological reports that set a diagnosis
of cancer or a laboratory report that evidences a diagnosis of HIV stated by the ELISA test. Patients cannot become
aware of this kind of information by reading it alone in their DMP and they need to be assisted by their doctor. Once
the medical consultation announcing the sensitive information has occurred, documents are no more sensitive, and they
become open, unless the patient decides to hide them.

The DMP is a personally controlled health record preserving patient rights: the patient should give her consent for a
health practitioner to create her DMP, she decides what can be seen and what must be hidden in her record, as well
as which health professionals are authorized to get access. The patient is provided with the journal of every access
to check for illegal accesses. The DMP also provides a personal section where the patient has the liability to create
various documents and record personal information she wants to share with her physicians.

Methods

After the publication in 2012 of the report from The Court of Auditors, the French Minister of Health ordered a study
to analyze the reasons for the poor adoption of the DMP and the solutions that could be implemented to improve
the situation. The question at this time was to know what was the best option between stopping the DMP project,
redefining a totally new project, or going on with the existing project but improving it. ASIP Santé conducted a nation
wide satisfaction survey with unstructured interviews and focus groups of patients and health professionals. Within
the Ministry of Health, the Department of Health Information Systems Strategy (DSSIS) (where one author, BS, is
working part-time as DMP project leader) was in charge of working with the different key stakeholders to recommend
the best strategy to adopt.

The method applied was to build a working group to gather the different stakeholders and have the different opinions
expressed regarding the DMP. We included all unions of general practitioners, the National Council of physicians, the
National Council of pharmacists, the French Federation of Medical Specialties, the French Association of Emergency
Physicians, ASIP Santé, the French National Authority for Health (HAS), the National federation of vendors (EHRs
for hospitals and private practices), patient organizations, chief information officers of main university hospitals. A
series of five meetings of the working group was organized in 2013 to discuss the future of the DMP project with a
diagnosis step (what went wrong?), and a “therapeutic” step to collectively construct a consensual roadmap.

All quantitative results are given on the basis of data provided by the ASIP Santé (DMP adoption data monthly sent to
the DSSIS) and the government health insurance.
Results

A Weak Adoption of the DMP. Consolidated data were obtained from December 2010, which is the date from which the DMP infrastructure was operational, until the end of December 2015. By this date, there were 567,888 DMPs at the national level. Figure 2 provides the distribution of DMP creations per week as well as the evolution of the number of DMPs on the study period. More than two thirds (72%) of all DMPs belong to patients older than 45 years (35% between 45 and 65, and 39% for older patients).

![Figure 2. Distribution of DMP creation per week and cumulative number of DMPs at the national level between December 2010 and December 2015.](image)

On the basis of the 38 million people namely insured, which are the potential DMP owners, 1.5% of the target population has a DMP after 5 years of deployment. As mentioned in the historical background, in 2009, ASIP Santé prospectively delivered yearly expectations of DMPs for the next five years reaching 13 millions DMPs in 2014.

As for DMP contents, 42% of DMPs are empty. The 330,273 non-empty DMPs contain as a whole 2,080,171 documents, with an average of 6.3 documents per DMP. These include the 129,826 (6.2%) documents that have been added by the patients themselves in less than 5% of all DMPs. Among all documents, 1,616 have been hidden by patients (less than 0.1%) and 36,501 were made sensitive by physicians (less than 2%). Among the 34 types of documents stored in DMPs, the most frequent is the report of medical consultation which represents 34% of all documents. Table 1 lists the 13 main types of documents (each type represents more than 1% of all documents). It must be noted that documents stored by patients come in fourth position representing 6.2% of all documents and that drug prescriptions are in eighth position with 2.9%.

A DMP user is recognized as such once she had created, feeded, or consulted at least one DMP. 690 healthcare structures, including public and private hospitals and clinics, are DMP users on an estimated total of about 75,000 healthcare structures. 6,636 private healthcare professionals, including mostly GPs, are DMP users. As a baseline, medical doctors in the private sector are about 130,000, in 2015 and other private healthcare professionals are 318,700 (less than 1%). Patients are DMP users when they logged at least one time into their DMP. They are 79,965 and represent 14.1% of all DMP owners and 24.2% of those for whom the DMP has been actually used and is not empty.

About one half of non-empty DMPs have been consulted (151,563), among which more than one half are also consulted by patients. Considered as a care coordination instrument, only 139,745 DMPs (42.3% of non empty DMPs) have been shared by 2 or more healthcare professionals.
Reasons for the weak adoption of the DMP. The fortune of national programs of HIT adoption varies according to the history of the countries. In France, unlike some countries, we were not starting from scratch, and it was much more difficult to change the components of the current system. From the different meetings of the working group gathered to elaborate the new DMP strategy, it appeared that the problem was cultural rather than technical although some technical aspects were not totally solved. Below are some reported issues:

No clear political support: there was no clear strong political support of the DMP program, and an incredibly noisy silence of political personalities on the subject. Because the DMP program crossed a lot of difficulties since its creation, successive Ministers of Health were not very keen on supporting it. After the report from The Court of Auditors in 2012, it took six months for the Minister of Health to say the DMP was not stopped, then six more months to announce the DMP2, the new version of the DMP, and another six months to advertise the program as part of the national health strategy. However, between these scarce interventions of the Minister of Health on the subject, there was no public information towards the different stakeholders. Suppliers of medical software stopped implementing DMP functionalities in their products. Health practitioners stopped investing in a tool which future seemed to be compromised, as illustrated by the decreasing rate of DMP creation (Figure 2).

No culture of medical information sharing among health practitioners: winning the bet of DMP adoption was relying on the assumption that doctors were ready and in demand of a tool to support the sharing of the medical information concerning their patients. But the patient-centered culture required to understand the benefits of sharing medical information was actually missing. Resistance came from health practitioners, especially from unions of physicians from private practice. If they could easily adhere to the concept of exchanging medical information, and were ready to move from paper-based letters to emails sent to known colleagues, they were reluctant to the concept of sharing, i.e. making relevant medical information available, for emergency and unscheduled care episodes, to practitioners unknown at the moment the information was delivered, but who could happen to manage the patient in the future. As opposed to the secure electronic messaging which is something acceptable since it is “just” replacing a known paper-based practice by a numeric exchange, the DMP program was a new object to support new practices, i.e. sharing medical information for the sake of care continuity. This object did not exist previously in “paper” version, and the organization and practices underlying its use did not exist in a formalized way. Thus, the adoption of both the concept and the tool was a challenge.

No education of patients and citizens: another difficulty came from the fact that patients were not actively involved in the DMP program. There was no national publicity campaign towards patients, giving some space to a negative press exacerbating the DMP program’s political sensitivity, associated with a wider “surveillance state”, the risk on data confidentiality, and the lost of the physician-patient privilege (in French “medical secret”). In addition, despite the “P” of DMP was for “personal”, patients could not create their DMP by themselves, and had to
ask their physicians to do it whereas most of them either were against the tool, or were thinking the program was stopped. In 2013, according to an opinion poll∗, 85% of French people were in favor of the DMP; an exceptionally high adherence rate (40% even said they were “strongly in favor”), in a marked progression in three years (they were 74% to be favorable in December 2010). Thus, involving patients and making them companions of the program would surely have helped the adoption of DMP beyond the benefit of empowering patients.

Persistence of technical difficulties: although 100% of EHR commercial software available in France are DMP-compatible, not all vendors provide integrated user-friendly DMP-specific functionalities. Besides, numerous health practitioners did not use the last DMP-compatible version of their medical software. Without incentive to change, because of the cost of the new software, and of the risks that always exist to modify a functional equilibrium, sometimes acquired with difficulties, health practitioners kept on using the old non DMP-compatible versions of their system, keeping the DMP outside their usual workflow, and making the adoption of the tool much more difficult.

Discussion

By the end of 2015, five years after the launching of the program in 2010, only 1.5% of potential DMPs had been created with a total of about 570,000 DMPs, and only 27% of all DMPs have been accessed. Updated data report 584,000 DMPs six month later, at the beginning of July 2016. As of February 2016, UK’s Health & Social Care Information Centre reported that 55 million people in England, representing more than 90% of the population, have an SCR created from their GP records†. According to the same source, 2.5 million SCRs were accessed by authorised healthcare professionals. Such difference in nationwide care records adoption between France and UK could be explained by the opt-out model used in the UK, the automated extraction of medical information from GPs’ EHRs to update the SCR, and the authority exercised by the NHS on GPs. Another important difference is that though being patient-centered, the SCR is not a “personal” record belonging to the patient who cannot acces it, but a professional, medical, tool to be used by health practitioners in case of emergency.11

The French DMP seems more comparable to the Australian personally-controlled electronic health record (PCEHR). Although not built on the same architectural model, the functionalities of the PCEHR22 are similar to those of the DMP, with an initial opt-in model, a richer document-based medical content, and the on-line patient access. The 2013-2014 annual report on the PCEHR23 mentions that, on June 30th, 2014, 1.73 million people registered for a PCEHR, i.e. 7.4% of the population two years after its launching. However, PCEHRs are mostly accessed by patients (around 512,000 times), and weakly accessed by health practitioners (around 25,000 times). The comparison with the US is more difficult due to the decentralized model for sharing medical information distributed among multiple EHRs through HIE.8 The number of people having a “personal record” does not make sense, since most patients have many records distributed within the medical practices they visit. However, not all providers are equipped with an EHR. The US CDC reported that in 2013, 22% of office-based physicians still did not use any type of EHR.24 Nevertheless, all patients may potentially access and download their medical information distributed among EHRs using the “Blue Button” functionality.25 This functionality builds upon HIE. From the patient’s point of view, this functionality is similar to the DMP and the PCEHR, except that the patient has no control over the content and cannot add or push any personal information. Recently, Ford et al.26 reported a simulation-based study to forecast the level of adoption of personal health records in the US. They estimate that the MU stage 3 target of 50% of consumers having access to their entire health record could be reached by 2020.

Conclusion

A nationally shared, widely accessible, electronic care record has a powerful symbolic meaning; it may be perceived as improving the quality and safety of care or, alternatively, as threatening patient data confidentiality or the traditional role of doctors and nurses.10,27,28 It is thus important to communicate about the nature and the aim of the tool. In

---

France, the low adoption rate of the DMP highlights the complex socio-technical challenges of implementing HIT. In the DMP case, the most important difficulty comes from the cultural change DMP was involving. However, despite the reported poor level of DMP adoption, patients and most healthcare professionals seem now to be ready to use it, although some unions of private practice physicians are still against the program because they feel that it is the role of the GP (and not of the DMP) to be the hub of care coordination, without considering that one valuable aspect of the DMP is for non-coordinated and emergency care.

Following the announcements of the French Minister of Health in 2013, the DMP has been relaunched (again), and a DMP2 positioned as a professional tool to support care continuity and coordination for the benefit of patients has been advertised. It took then two years to publish the new French Health Act† (January 26th, 2016) that includes the promotion of DMP adoption and use (article 96). After considering that the name of the tool should be changed, it remained DMP but with “P” for “partagé” (shared) instead of “personal” to insist on its use to support care continuity among health professionals sharing the same relevant medical information. Instead on depending on busy doctors reluctant to perform administrative tasks, patients will be now able to create their own DMP which also helps to solve the problem of tracing the consent collection. Instead of ASIP Santé, it is now the responsibility of the government health insurance to conduct the DMP program. Indeed the poor DMP adoption has been partly explained by a principled opposition of private doctors and by some technical difficulties private doctors may face in updating their EHR system. Thus having the government health insurance that financially monitors private doctors and is already used to send computer technicians to doctors’ offices to check whether computerized billing tools are operational was a better candidate. In addition, the government health insurance represents a significant counterpower towards the medical software vendors, which should also improve their reactivity to make the appropriate technical developments for interfacing their medical software with the DMP functionalities. Another important point is that the government may push all health insurance claims information (including information about reimbursed medical consultations, reimbursed drugs, reimbursed biological exams, imaging prescriptions, among other clinical information), thus ensuring there would be no more empty DMPs. Finally, the government health insurance may use the ROSP program to offer financial incentives and make private doctors actually use (create, feed and read) the DMP in a kind of French Meaningful Use. The impact of this year’s new DMP initiative will have to be assessed and quantified.

Acknowledgments

Authors thank Dr Elie Lobel, from ASIP Santé, for the information he provided about the DMP program, as well as Philippe Burnel and Philippe Cirre from the DSSIS at the French Ministry of Health. Data and figures on DMP usage have been gracefully provided by ASIP Santé.

References


---

†https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT0000031912641 [accessed July 7th, 2016].


Identification and Use of Frailty Indicators from Text to Examine Associations with Clinical Outcomes Among Patients with Heart Failure

Yijun Shao, Ph.D. 1, April F. Mohanty, Ph.D., M.P.H. 2, Ali Ahmed, M.D., M.P.H.3, Charlene R. Weir, Ph.D., R.N. 1,2, Bruce E. Bray, M.D. 1, Rashmee U. Shah, M.D., M.S.4, Douglas Redd, Ph.D.1,2,3,5, Qing Zeng-Treitler, Ph.D.1,2,3,5

1Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, USA; 2Informatics Decision-Enhancement and Analytic Sciences (IDEAS) Center, George E. Wahlen Department; Veterans Affairs Medical Center, Salt Lake City, UT, USA; 3Washington DC VA Medical Center, Washington, DC, USA; 4University of Utah, Division of Cardiovascular Medicine, Salt Lake City, UT, USA; 5Department of Clinical Research and Leadership, George Washington University, Washington, DC, USA

Abstract

Frailty is an important health outcomes indicator and valuable for guiding healthcare decisions in older adults, but is rarely collected in a quantitative, systematic fashion in routine healthcare. Using a cohort of 12,000 Veterans with heart failure, we investigated the feasibility of topic modeling to identify frailty topics in clinical notes. Topics were generated through unsupervised learning and then manually reviewed by an expert. A total of 53 frailty topics were identified from 100,000 notes. We further examined associations of frailty with age-, sex-, and Charlson Comorbidity Index-adjusted 1-year hospitalizations and mortality (composite outcome) using logistic regression. Frailty (≥ 4 topics versus <4) was associated with twice the risk of the composite outcome, Odds Ratio: 2.2, 95% Confidence Interval: (2.0-2.4). This study demonstrates the feasibility of identifying frailty indicators from clinical notes and linking these to clinically relevant outcomes. Future work includes integrating frailty indicators into validated predictive tools.

Introduction

Frailty is an important, but frequently overlooked, determinant and indicator of health outcomes in older adults1-4. It is distinct from comorbidity and disease, and is an age-related, multidimensional combination of fatigue, weakness, malnutrition, and greater vulnerability to stressors5-7. Increasing frailty is often co-existent among patients diagnosed with chronic conditions and it is associated with the severity of acute illness and intensity of healthcare utilization8, 9. Frailty is widely recognized as a risk factor for morbidity, including cognitive decline, mortality, as well as a barometer of how well patients respond to treatment5, 9. Therefore, this metric can be valuable for informing healthcare decisions, including recommended surgical procedures, end-of-life care, and living arrangements1. In order to provide an efficient and effective patient-centered care, providers should be aware of each patient’s frailty status for incorporation into clinical decision making.

Despite its importance, frailty measurements are rarely collected in a quantitative, systematic fashion in routine patient care9. Frailty extraction is complex, since information is rarely collected as coded/structured data and clinical note terminology varies significantly. This complexity limits the ability to evaluate the association between frailty severity and clinical outcomes, and prevents healthcare providers from using frailty information in clinical decisions9. However, healthcare providers commonly document various aspects of frailty in clinical notes, especially when treatment plans change and/or when patients’ quality of life is impacted. Therefore, clinical notes can provide the data for a quantitative frailty metric as well as a longitudinal record of frailty.

Extracting frailty from clinical notes is challenging because frailty is a complex and multi-faceted concept. The use of the term “frail” is not common, while the description of frailty indicators such as “tired,” “slow walk,” “unsteady gait” and “confused” are abundant. In a prior study, we identified terms describing functional status, one aspect of frailty10. Among the hundreds of functional status terms we collected from clinical notes and social media, only a fraction could be mapped to a controlled vocabulary. Furthermore, for the purpose of retrospective analysis or prospective decision support, frailty should not be treated simplistically as a binary variable. The Frailty Scale developed by Dr. Rockwood, for instance, categorized frailty status into 9 levels11. As a result, to determine frailty

1110
status from text, we cannot rely on a small number of known terms or the existing controlled vocabulary. Rather, it is necessary for us to identify a rich set of indicators of frailty and differentiate the degrees of frailty.

Topic modeling is a potential approach for identifying indicators of frailty in clinical notes. Briefly, topic modeling is a machine learning method based on statistical models that assess words in documents for identifying hidden themes in a corpus of free text documents\textsuperscript{12}. By first focusing on themes rather than individual words/terms, we hypothesize that it would allow us to capture a wide range of text descriptions potentially related to frailty.

Our aims are to 1) demonstrate the feasibility of using topic modeling of clinical notes to identify frailty indicators and 2) using these frailty indicators to examine associations with clinical outcomes among patients diagnosed with heart failure (HF). HF is a highly prevalent condition that results from structural or functional impairment of the heart’s ability to fill or eject blood. Recent projections estimate that >10 million Americans (3%) will have HF by 2030\textsuperscript{13,14}. Evidence suggests that frailty is worse among patients with HF compared to similarly aged adults, with similar comorbidities, from the general population\textsuperscript{15-17}. Commonly reported HF symptoms such as dyspnea and fatigue can lead to greater frailty\textsuperscript{14} and one of the treatment goals is to improve frailty status. Also, since HF patients may be referred for cardiac surgery or other invasive interventions as symptoms worsen (though less common among older HF patients), measurement of frailty can inform treatment strategies and facilitate examination of treatment response\textsuperscript{14}. A method to systematically estimate frailty among all HF patients is, therefore, critically important.

**Methods**

**Data**

We used Veterans’ electronic medical record (EMR) data which are maintained by the Veterans Administration Informatics and Computing Infrastructure (VINCI) database\textsuperscript{18}. VINCI includes data on healthcare utilization (inpatient and outpatient encounters), clinical parameters, and demographics (e.g. age, gender) as well as narrative text from clinical encounters or Text Information Utility (TIU) notes. We used the International Classification of Disease 9th Clinical Modification (ICD-9-CM) codes 428.0-428.9 from inpatient (primary/principal or secondary diagnoses) and outpatient records to identify patients with HF. We identified 12,000 patients with a HF diagnosis in Fiscal Year 2010, among which 4,000 experienced death or ≥2 HF hospitalizations during the year after diagnosis. Then we extracted Veterans’ gender, calculated their age and Charlson Comorbidity Index (CCI) at the first HF diagnosis. The Charlson Comorbidity Index is a measure of disease burden that is determined from comorbid conditions using methods described by Quan et al.\textsuperscript{19}. These Veteran characteristics were either directly or indirectly ascertained from the structured data tables. We also retrieved all the clinical notes from the patients included in our study dated within one year before the first HF diagnosis. There were a total of 709,389 notes available for the 12,000 Veterans diagnosed with HF. Frailty indicators, which were not available from the structured data, were extracted from these notes using a topic modeling technique. All data access and use was approved through local Institutional Review Board and VINCI policies and no human subjects were contacted for this study.

**Topic Modeling and Stable Topic Extraction**

We randomly sampled 100,000 notes from the corpus of 709,389 notes for topic modeling. Random samples (50,000 each) were balanced between the 4,000 patients with and the 8,000 patient without 1-year outcomes (≥2 HF hospitalizations or mortality). We tokenized these texts by converting all uppercase letters into lowercase, removing all numbers and punctuations except hyphens (“-”) that join two words. We used the Latent Dirichlet Allocation (LDA)\textsuperscript{20} program from the MALLET package, a topic modeling tool that is written in Java\textsuperscript{22}. LDA is one of the most widely used topic modeling approaches and makes assumptions that topics are probabilistic distributions over words and documents are mixtures of topics. It has found applications in many areas including biomedicine\textsuperscript{12,21}. We set the initial number of topics to be 700 and ran LDA on the 100,000 notes to obtain a list of 700 topics. Then we applied the learned topic model to the full collection of 709,389 notes using the topic inferring tool also included in MALLET. This step yields topic proportions per note for each note. By the design of LDA, every topic has a non-zero proportion within every note. We set 0.02 as the cut-off on the topic proportion in a note to select the topics present in the note.

One major issue when using LDA is the presence of a high amount of noise (irrelevant terms) in some topics, which results in high false positives. To overcome this problem, we designed a method to identify the non-noise topics which we called “stable” topics. Since LDA is an iterative algorithm that initially assigns terms randomly to topics, the topic model it generates varies each time LDA is applied to a test corpus. At the same time, certain topics re-
occur in repeated experiments. For example, in this study we observed the following two topics (only top 10 most probable words are listed here) from two independent runs:

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>walker</td>
<td>1440</td>
<td>0.106</td>
</tr>
<tr>
<td>gait</td>
<td>1071</td>
<td>0.079</td>
</tr>
<tr>
<td>cane</td>
<td>974</td>
<td>0.072</td>
</tr>
<tr>
<td>ambulates</td>
<td>433</td>
<td>0.032</td>
</tr>
<tr>
<td>steady</td>
<td>433</td>
<td>0.032</td>
</tr>
<tr>
<td>ambulating</td>
<td>429</td>
<td>0.032</td>
</tr>
<tr>
<td>ambulation</td>
<td>421</td>
<td>0.031</td>
</tr>
<tr>
<td>walking</td>
<td>320</td>
<td>0.024</td>
</tr>
<tr>
<td>unsteady</td>
<td>312</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Run 1, Topic 24

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>walker</td>
<td>1799</td>
<td>0.114</td>
</tr>
<tr>
<td>gait</td>
<td>1301</td>
<td>0.083</td>
</tr>
<tr>
<td>cane</td>
<td>1172</td>
<td>0.074</td>
</tr>
<tr>
<td>ambulation</td>
<td>467</td>
<td>0.030</td>
</tr>
<tr>
<td>walking</td>
<td>437</td>
<td>0.028</td>
</tr>
<tr>
<td>balance</td>
<td>407</td>
<td>0.026</td>
</tr>
<tr>
<td>steady</td>
<td>388</td>
<td>0.025</td>
</tr>
<tr>
<td>ambulates</td>
<td>386</td>
<td>0.025</td>
</tr>
<tr>
<td>unsteady</td>
<td>334</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Run 2, Topic 144

Although the topics do not have exactly the same probability distributions over the words, they are very similar, and should be viewed as the “same” topic that recur in different runs.

We postulate that non-recurring topics and topics terms are likely to be noise as they are influenced by the initial random seeds. To identify the recurring topics, we proceeded as follows. Given two topics \(a\) and \(b\) with word counts \(c_t^a\) and \(c_t^b\) respectively, we define the distance between \(a\) and \(b\) to be

\[
d(a, b) = 1 - \frac{\sum_w c_t^a(w) \times c_t^b(w)}{\sqrt{\sum_w (c_t^a(w))^2} \times \sqrt{\sum_w (c_t^b(w))^2}}
\]

which is known as the cosine-distance. To identify stable topics generated by LDA corresponding to some parameter set (e.g., number of topics, number of iterations, alpha), we ran LDA 3 times under the same parameter set (but with different random seeds) to produce 3 sets of topics, say, \(A, B, C\). Then, for each topic triple \((a, b, c)\) with \(a \in A, b \in B, c \in C\) we compute the size of the triple by taking the maximum of the 3 distances:

\[
d(a, b, c) = \max \{d(a, b), d(a, c), d(b, c)\}
\]

One can visualize a topic triple as a triangle, so that the terms such as “size”, “small”, “large” make more sense.

We aligned up the three sets of topics using a greedy algorithm as follows. We first found the triple \((a_1, b_1, c_1)\) of the smallest size from all such triples, and second found the triple \((a_2, b_2, c_2)\) of the smallest size from all the triples \((a, b, c)\) with \(a \in A \setminus \{a_1\}, b \in B \setminus \{b_1\}, c \in C \setminus \{c_1\}\), and third found the triple \((a_3, b_3, c_3)\) of the smallest size from all the triples \((a, b, c)\) with \(a \in A \setminus \{a_1, a_2\}, b \in B \setminus \{b_1, b_2\}, c \in C \setminus \{c_1, c_2\}\), and so on. Thus we aligned the three sets of topics by their size from small to large. For any topic triple \((a, b, c)\) smaller than an adjustable parameter \(r\) we made a new topic by averaging the counts of \(a, b\) and \(c\):

\[
c_{t_{avg}}(w) = \frac{1}{3}[c_t^a(w) + c_t^b(w) + c_t^c(w)]
\]

and we called the new topic a stable topic. In this study, we empirically set \(r = 0.7\) and obtained 556 stable topics.

To determine what stable topics were present in each note, we proceeded as follows. First we independently applied each of the 3 learned LDA models to the full collection of 709,389 notes using the topic inference tool included in MALLET. This step yielded 3 topic proportions per note for each note. Next, we defined a stable topic to be present in a note if at least 2 of the 3 topics in the topic triple corresponding to the stable topic had a proportion of 0.02 or higher in that note. This enabled us to make a list of stable topics present in a note for each note.
Topic Interpretation and Frailty Topic Identification

The “topics” generated by topic modeling methods are simply probability distributions over the whole vocabulary (consisting of thousands of words), and every word has a non-zero probability in every topic. Since we need to identify topics related to frailty, the first step is to interpret all the probability distributions over words. Obviously, not all the words are equal for making interpretations: words with higher probabilities are more important than those with lower probabilities. For our project, we list the top 20 most probable words for each topic and ask the human reviewers to focus on the top 10 words to make interpretation and use the next 10 words as auxiliary.

An informatics expert and a medical expert independently reviewed the list of 556 stable topics, interpreted them as described above, and identified those that represent certain aspect(s) of frailty. Then we calculated the inter-rater agreement on the identification of frailty topics. See Table 1 for examples of frailty topics and their interpretations.

Counting Frailty Topics for Patients

We counted the frailty topics for each patient from all the notes dated within one year of the first diagnosis of HF. The topics were counted only once, even if they appeared in multiple notes. These numbers were recorded as the frailty data. We hypothesized that the number of frailty topics a patient has can be used as an indicator of the frailty severity: the higher the number, the more severe the frailty.

Cases and Controls

We divided the 12,000 HF patients into case and control groups. Cases were defined as patients who died or had \( \geq 2 \) all-cause hospitalizations within the first year following their first diagnosis of HF (our composite outcome). Controls were patients who did not have the composite outcome within the 1-year period following their initial HF diagnosis. This yielded 5,207 cases and 6,793 controls.

Statistical Analyses

To explore the relevance of the frailty indicators, we ascertained the composite outcomes (mortality/hospitalization) and we conducted several statistical analyses.

1) Correlation Analysis

We computed the prevalence of patients with \( n (n=0, 1, 2, \ldots) \) frailty topics in the cases and also in controls. Then we calculated ratio of the prevalence in cases to those in controls for each \( n \). These ratios measure the population level (i.e., cases vs controls) difference in frailty of various levels of severity. Therefore, the correlation between the ratios and the number of frailty topics (i.e., \( n \)) is an indicator of the relationship between frailty and the composite outcome.

2) Logistic Regression Models

We used logistic regression models to analyze the association of frailty and the composite outcome at the individual level. We adjusted for age, gender, and CCI level. The outcome was modeled as a binary variable: case was coded as 1 and control was coded as 0 (reference). Age (in years) was measured at the time of the first HF diagnosis. CCI was an integer ranging from 0 to 33. Age and CCI were modeled as continuous variables. Gender was modeled as a binary variable: 0 for male and 1 for female. We parameterized the frailty variable in two ways: 1) as a continuous variable and 2) as a binary variable. As a continuous variable, frailty was coded as the number of frailty topics we counted for each patient. As a binary variable, frailty was coded as 0 (reference) if the number of frailty topics was \( \geq 4 \) topics and coded as 1 for \(<4\) topics. We chose 4 as the cut-off based on the results of correlation analysis (see Figure 1 and 2). For both cases, we calculated the odds ratio for the composite outcome associated with the frailty variable based on the coefficients output by the logistic regression models.

Results

We obtained 556 stable topics from the topic modeling step. An informatics expert reviewed these topics and identified 53 of them as frailty topics. A medical expert independently identified frailty topics resulted and had an inter-rater agreement of Kappa = 0.818 with the first reviewer, demonstrating that the first expert’s result was reliable. In Table 1, we list eight frailty topics for illustration.

The age, gender, and CCI characteristics of our study population of 12,000 Veterans with new diagnoses of HF are described in Table 2. Mean age was around 70 years old for cases and controls, most were male, and cases generally had a higher CCI compared to controls.
Table 1. Eight examples of frailty topics. The first column lists the labels assigned by the informatics expert as human interpretations. The second column lists the top 20 words in order of descending probability.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top 20 most probable words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall Risk</td>
<td>fall risk score morse gait ambulatory status scale mental falling secondary diagnosis total</td>
</tr>
<tr>
<td></td>
<td>oriented history aid ability iv lock bedrest</td>
</tr>
<tr>
<td>Bedside</td>
<td>bed resting monitor noted distress continue complaints quietly sleeping bedside discomfort</td>
</tr>
<tr>
<td>Condition</td>
<td>comfortably watching tv lying remains vss received unlabored sitting</td>
</tr>
<tr>
<td>Bedside</td>
<td>bed call reach light position fall low side locked rails safety assistance lowest room free</td>
</tr>
<tr>
<td>Assist</td>
<td>risk precautions environment monitor place</td>
</tr>
<tr>
<td>Home Care</td>
<td>home care services health va referral agency community nursing service aide hospice hha</td>
</tr>
<tr>
<td></td>
<td>skilled referred fax visits medicare fee start</td>
</tr>
<tr>
<td>Family</td>
<td>daughter son home called spoke father call family states law dtr living told stated lives</td>
</tr>
<tr>
<td>Assist</td>
<td>assisted daughters dementia received aware</td>
</tr>
<tr>
<td>Mobility</td>
<td>walker cane gait wheelchair ambulation steady ambulates ambulating walking unsteady</td>
</tr>
<tr>
<td>Assist</td>
<td>equipment device home ambulate falls walk assistance balance scooter mobility</td>
</tr>
<tr>
<td>Cognitive</td>
<td>memory cognitive average speech evaluation range difficulty recall words communication</td>
</tr>
<tr>
<td>Status</td>
<td>functioning task attention dementia information impaired ability deficits accuracy impairment</td>
</tr>
<tr>
<td>Shoe-wear</td>
<td>shoes wear issued fit size instructed wearing prosthetics proper device equipment care</td>
</tr>
<tr>
<td>Ability</td>
<td>stockings compression shoe hose wheelchair scooter measured order</td>
</tr>
</tbody>
</table>

Table 2. Study population characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Cases N = 5,207</th>
<th>Controls N = 6,793</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (year, mean(SD))</td>
<td>70.5 (13.0)</td>
<td>69.1 (11.4)</td>
</tr>
<tr>
<td>Men (%)</td>
<td>5096 (97.9)</td>
<td>6647 (97.9)</td>
</tr>
<tr>
<td>Charlson Comorbidity Index (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1248 (24.0)</td>
<td>2516 (37.0)</td>
</tr>
<tr>
<td>1</td>
<td>953 (18.3)</td>
<td>1451 (21.4)</td>
</tr>
<tr>
<td>2</td>
<td>722 (13.9)</td>
<td>998 (14.7)</td>
</tr>
<tr>
<td>3</td>
<td>661 (12.7)</td>
<td>717 (10.6)</td>
</tr>
<tr>
<td>4</td>
<td>543 (10.4)</td>
<td>515 (7.6)</td>
</tr>
<tr>
<td>5+</td>
<td>1080 (20.7)</td>
<td>596 (8.8)</td>
</tr>
</tbody>
</table>

Among cases, 20.3% patients had no frailty topics mentioned in the notes, 8.7% had one frailty topic, 8.4% had two frailty topics, 6.5% had three, 6.5% had four, 5.5% had five, etc. Among controls, 35.4% patients had no frailty topics, 13.4% had one, 10.8% had two, 8.8% had three, 6.1% had four, 4.8% had five, etc.

Comparing the two distributions, when the number of frailty topics was less than 4, the prevalence of patients having that number of topics in the cases was lower than that in controls. However, patients with ≥ 4 topics, the prevalence of patients having that number of topics was higher in the cases than in the controls. This is illustrated in Figure 1. This led to the choice of 4 as the cut-off in making frailty into a binary variable in one of logistic regression models.

To make the contrast clearer, we calculated the ratio of the prevalence in cases to that in controls for each number of topics, except for 25 or more topics where we grouped them together into one category. The correlation of this ratio with the number of frailty topics is 0.879.
Figure 1. Patient distribution over different ranges of numbers of topics in cases and in controls. The ranges are: 1) 0 topics, 2) 1 ~ 3 topics, 3) 4 ~ 10 topics, 4) 11 ~ 24 topics, 5) 25 or more topics.

Figure 2. Ratio of prevalence of patients with various numbers of frailty topics in the cases to that in the controls. Correlation of this ratio to the number of frailty topics was 0.879.

Figure 2 indicates that the ratio was below 1 when the number of frailty topics (n) was below 4, between 1 and 2 when n was between 4 and 10, and above 2 when n was above 11 (the only exception was when n=20 should be viewed as an effect of random sampling).

The results of logistic regression models are presented in Tables 3 and 4, corresponding to treating frailty as a continuous variable and a binary variable (reference, <4 topics), respectively. In both models, frailty was associated with the outcomes and these results were statistically significant. The coefficient (0.07) of frailty in Table 3 can be interpreted as the odds ratio = 1.07 (=exp(0.07)), 95% Confidence Interval (CI): 1.07-1.08, for each additional frailty topic associated with the risk of the composite outcome. Thus, each additional frailty topic was associated with a 7% higher risk of the composite outcome. The coefficient (0.789) corresponds to an odds ratio of 2.20 (95% CI: 2.02-2.39) for the association between ≥4 frailty topics (vs <4) and the composite outcome, Table 4. In other words, ≥4
frailty topics was associated with two times the risk of having the composite outcome 1-year following an initial HF diagnosis.

**Table 3. Logistic regression with frailty modeled as a continuous variable (reference, 0 topics)**

<table>
<thead>
<tr>
<th>Intercept</th>
<th>Age</th>
<th>Gender</th>
<th>CCI</th>
<th>Frailty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>-1.371</td>
<td>0.008</td>
<td>-0.127</td>
<td>0.099</td>
</tr>
<tr>
<td>95% CI</td>
<td>(-1.595, -1.146)</td>
<td>(0.004, 0.011)</td>
<td>(-0.391, 0.138)</td>
<td>(0.080, 0.118)</td>
</tr>
<tr>
<td>P-value</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.347</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Table 4. Logistic regression with frailty as a binary variable (reference, <4 topics)**

<table>
<thead>
<tr>
<th>Intercept</th>
<th>Age</th>
<th>Gender</th>
<th>CCI</th>
<th>Frailty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>-1.399</td>
<td>0.008</td>
<td>-0.069</td>
<td>0.122</td>
</tr>
<tr>
<td>95% CI</td>
<td>(-1.622, -1.175)</td>
<td>(0.004, 0.011)</td>
<td>(-0.328, 0.191)</td>
<td>(0.104, 0.140)</td>
</tr>
<tr>
<td>P-value</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.604</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Discussion and Conclusion**

In this study we demonstrate the feasibility of extracting frailty-related topics from clinical text and that these topics are associated with clinically relevant outcomes. We are unaware of previous studies that have extracted frailty-related topics from clinical notes and associated these topics for clinical outcomes. Our results show that the frailty data we extracted is correlated with mortality and/or hospitalizations at both the population and the individual levels. In particular, the number of frailty topics for a given patient may serve as a proxy for the severity of frailty of the patient. A patient with one or two frailty topics does not necessarily mean the patient is frail: for example, that a patient needs mobility assistance may be a result of injury. However, with more frailty topics, it is more likely that the patient is frail.

Our results have the potential for significant impact. Healthcare providers use risk prediction models to estimate treatment risk. For example, the Society of Thoracic Surgeons uses a calculator to estimate the risk of adverse clinical outcomes following cardiac surgeries. This calculator is inaccurate for older adults—it underestimates risk in frail patients and overestimates risk in healthy patients. A tool that allows frailty to be incorporated into the STS, and other, models could improve the accuracy of risk prediction. We hope that our project can lead to a better measure of frailty that can be useful in routine use of clinical decision support.

Our findings are consistent with previous studies of frailty, not just indicators of frailty, and clinical outcomes. For instance three of the most commonly used measures of frailty include the Fried Frailty Phenotype model, the Frailty Index developed by Rockwood et al., and gait speed. These measures have been found to be associated with future risk of hospitalizations and mortality. However, these measures are not integrated in routine care for all patients, since they can be time consuming to complete and add additional burden to patients and providers. Thus, these measures are often reserved for randomized controlled studies, as in the recent landmark Systolic Blood Pressure Intervention Trial (SPRINT), or smaller target populations, such as among nursing home residents. Thus, the potential use of clinic notes from the EMR supports opportunities for larger and more generalizable or pragmatic studies of frailty and incorporation into routine healthcare decision making.

We can conclude that the topic modeling approach is capable of extracting the frailty information from text. This approach is also applicable to much larger datasets. Because LDA is a time-consuming program, instead of running LDA directly on all the notes (about 700,000 notes), we sampled a sufficiently large subset (100,000 notes) so that LDA can finish the job in a reasonable time. Then we applied the learned topic model to the all the notes to infer the topic distributions in each of them. The topic inferring program is much less time-consuming than LDA, so the process can finish in a reasonable time while still obtaining the outputs we need: the list of topics and the topic distributions in each note. One novelty of our topic modeling approach is the stable topic identification method.

Topic modeling is an unsupervised machine learning method, which means it does not require human annotations to train the model. Therefore our approach does not require much labor-intensive work such as reading notes and manual annotation. That being said, some manual work – interpreting the computer-extracted topics and identify those that could be indicators of frailty – is still required. However, the human labor is much less than annotating notes, because a reviewer only needs to read 20 words (in this study) for a topic rather than hundreds of words for a note.
Some frailty topics identified by our experts may appear to be cognition-related (see Table 1). This is because there is an association between frailty and cognitive impairment: frailty and cognition interact within a cycle of age-associated decline, therefore some researchers regard cognition deficits as indicators of frailty.

Since this study is our first attempt of identifying frailty from text, there are a number of limitations. 1) A note having a frailty topic could indicate absence of frailty. Using a number of (e.g. ≥ 4) distinct frailty topics as indicator greatly increases the chance of finding frailty, although the certainty is still not 100%. 2) When we used the number of frailty topics as an indicator of the severity of frailty, we implicitly assumed that all the frailty topics had the same contribution to the severity of frailty. In reality different frailty topics may indicate different levels of severity. 3) Our logistic regression analysis only took into account 3 commonly used predictors in HF patients for our composite outcome: age, gender and CCI. It is likely that many other factors that also correlate to the outcome of mortality/hospitalizations. 4) Although we attempted to identify Veterans with a new or incident diagnosis of HF, patients could have sought care outside the Veterans Health Administration for their initial diagnosis and would have been included in this study of incident HF patients. In addition to addressing these limitations, future studies are needed to replicate our preliminary findings and to extend our findings to other clinical domains. Also, further work is needed to examine changes in frailty indicators from clinical notes over time and to incorporate frailty indicators into validated predictive tools for clinical decision making.

Based on the results of this study, we conclude that it is feasible to identify indicator of frailty from routinely documented clinical notes. We also demonstrate that these indicators are related to clinically relevant outcomes in a population of Veterans diagnosed with HF.

Acknowledgements

This work is funded by the US Department of Veterans Affairs, Office of Research and Development, Health Services Research and Development grants CHIR HIR 08-374, HIR 08-204, CRE 12-315 and the CREATE: A VHA NLP Software Ecosystem for Collaborative Development and Integration. Dr. Mohanty is supported by the VA Advanced Fellowship Program in Medical Informatics of the Office of Academic Affiliations, Department of Veterans Affairs. We would also like to acknowledge the staff, resources and facilities of the VA Salt Lake City IDEAS Center.

References


Standardized Representation of Clinical Study Data Dictionaries with CIMI Archetypes

Deepak K. Sharma\(^1\), Harold R. Solbrig\(^1\), Eric Prud'hommeaux\(^2\), Jyotishman Pathak\(^3\), Guoqian Jiang\(^1\)

\(^1\) Mayo Clinic College of Medicine, Rochester, MN; \(^2\) W3C/MIT, Boston, MA; \(^3\) Weill Cornell Medical College, New York, NY

Abstract

Researchers commonly use a tabular format to describe and represent clinical study data. The lack of standardization of data dictionary’s metadata elements presents challenges for their harmonization for similar studies and impedes interoperability outside the local context. We propose that representing data dictionaries in the form of standardized archetypes can help to overcome this problem. The Archetype Modeling Language (AML) as developed by the Clinical Information Modeling Initiative (CIMI) can serve as a common format for the representation of data dictionary models. We mapped three different data dictionaries (identified from dbGAP, PheKB and TCGA) onto AML archetypes by aligning dictionary variable definitions with the AML archetype elements. The near complete alignment of data dictionaries helped map them into valid AML models that captured all data dictionary model metadata. The outcome of the work would help subject matter experts harmonize data models for quality, semantic interoperability and better downstream data integration.

1 Introduction

Spreadsheets and other tabular formats have become the primary mechanism for representing much of the clinical research data that is published today. The simple format and easy-to-use interface has allowed researchers to customize and manage data with relative ease. Large projects like the Database of Genotypes and Phenotypes (dbGaP)\(^1\),\(^2\), the Phenotype Knowledge Base (PheKB)\(^3\),\(^4\) and The Cancer Genome Atlas (TCGA)\(^5\),\(^6\) have published their data dictionaries in a variety of formats. They have used either delimited values variants (CSV - comma-separated values) or Extended Markup Language (XML) schemas, where column headings or XML elements (and associated attributes) carry an implicit metadata about the study and its variables. While a simple representation makes it easier to procure, maintain and disseminate datasets and implementation artifacts in these formats, it presents a challenge to the semantic interoperability of the models that these data dictionaries define. These model assertions, even when they are standardized with external vocabulary resources, almost always require additional transformations to be effective and semantically interoperable outside their local context.

There is an emerging need for standard representation of these data dictionaries to facilitate their harmonization and for greater semantic interoperability\(^7\),\(^11\). A standard representation for a data dictionary would allow researchers to identify the minimal set of information that would be needed to align their model with other models that describe same set of semantics. This would also enhance the clarity of the individual model semantics and improve their interoperability when model and its datasets are shared. A standard template for the representation of a data dictionary could greatly reduce the transformations required to move between different environments.

The Archetype Modeling Language (AML)\(^12\) is one of formalisms approved by the HL7 Clinical Information Modeling Initiative (CIMI)\(^13\),\(^14\) to describe clinical models. An AML model is represented using the Unified Modeling Language (UML)\(^15\). The AML specification is based on elements from HL7 Detailed Clinical Models (DCM)\(^16\), OpenEHR ADL Archetypes\(^17\),\(^18\), the ISO 11179 Meta Data Repository (MDR)\(^19\) and the OMG Common Terminology Services 2 (CTS2)\(^20\) standards. AML archetypes employ constraint-based modeling approach, where the possible values of a starting reference model are incrementally restricted (constrained) as the target of the model is specialized.

The objective of the study is to create direct mappings between the metadata from a collection of three data dictionaries (from dbGAP, PheKB, and TCGA, respectively) with their variable definitions and the AML constraints on the CIMI Core Reference archetypes\(^21\), and demonstrate the utility of the mappings by implementing the transformation via a suite of semi-automated tools. The treatment of data dictionary variable definitions as AML constraints provides both a standardized representational form and facilitates their harmonization with other similar archetypes. Based on the mappings, we develop a platform known as D2Refine on top of an open-source...
OpenRefine platform. The platform provides a suitable interface that allows users to load data dictionaries in a variety of formats, align them to a standard template, and transform them into CIMI archetypes.

2 Materials and Methods

2.1 Materials

2.1.1 Clinical Study Data Dictionaries

We used three data dictionary formats identified from dbGaP, PheKB and TCGA. The data dictionaries from dbGaP and TCGA are in XML schema format whereas the format of PheKB is Excel spreadsheet. Each of the data dictionaries contains a collection of metadata including variable definition, constraints and links to standard terminologies. Some of these data dictionaries include references to standard terminologies like NCI Thesaurus and Common Data Elements from Cancer Data Standard Registry and Repository (caDSR) already embedded in most of the models they define.

2.1.2 OMG AML Specification

The AML specification, as an Object Management Group (OMG) standard, is composed of three UML Profiles - Reference Model Profile (RMP), Terminology Profile (TP) and Constraint Profile (CP). The Reference Model Profile defines stereotypes to identify the ground rules for constraining a target UML model. The Terminology Profile provides a set of stereotypes that all UML models to have multi-lingual names and comments, allows the assignment of multiple external identifiers (e.g. URIs etc.) and allows UML model elements to be associated with their intended semantics through links to elements from external ontologies. The Constraint Profile defines a set of stereotypes that allows the specification of constraints (archetypes) on the target classes, properties and data types in a target UML model. An AML archetype is a set of constraints on a target class in a target reference model. An AML archetype library is a collection of archetypes that constrain one or more classes in a given UML model (the Reference Model or RM). Note that additional AML documentation includes 1) the AML specification in GitHub (including a collection of normative and non-normative formats of AML artifacts), 2) the AML Archetype Examples, 3) the OpenEHR’s ADL and AOM Specifications and 4) an AML prototype project in GitHub.

2.1.3 The Reference Model

CIMI has specified a generic reference model, the CIMI Core Reference Model that underlies all archetypes published by the CIMI group. CIMI also publishes a set of “reference archetypes”, the CIMI Core Reference Archetypes. These archetypes constrain the CIMI Core Reference Model and form the basis of all other archetypes published by the CIMI group. In the clinical space, these reference archetypes provide basic structures to catalog the majority of the constraint definitions listed in data dictionaries. Figure 1 shows a selection of CIMI Core RM classes and a high level CIMI Core Reference Archetype 'Cluster' (an online version available at *).
2.1.4 OpenRefine - The Platform

We will use an open-source platform known as OpenRefine\textsuperscript{22}, which is a popular and widely used tool to manage and clean up data. OpenRefine provides a mechanism that allows users to programmatically extend its capabilities to add standard templates in a tabular form. The standard template enables users to define mandatory and optional variable definitions for a data dictionary. The OpenRefine’s built-in reconciliation feature (which is extensible) allows users to link the values with external terminology services. The set of constraints in a template can then be programmatically transformed into CIMI archetypes by extending the export/import functionality of OpenRefine. We plan to add these features without altering spreadsheet-like interface of OpenRefine intact. This greatly minimizes the learning curve for users.

2.2 Method

2.2.1 Create a Standard Template based on AML specification

The first step towards capturing the archetype requirements is the creation of a Standard Template. The Standard Template serves as a bridge between tabular metadata and their AML equivalents. The template provides a minimum set of requirements to be filled by variable definitions of a data dictionary and its metadata. The Standard template would need to be implemented once and hence it would eliminate the need for any additional transforms. For a data dictionary, what we need to do is to map it to the Standard Template and to get it transformed into an AML archetype.

The elements in the Standard Template are, directly derived from the AML specification profiles and are manually composed, as a set of three types of requirements to create a valid AML archetype and to link its terms to standard terminologies. These requirements are:

1. Archetype Metadata: archetype library, identification, constrained RM class, archetype specialization, organization, copyright, version and author information
2. Constraint Definition: constraint identification, constrained type, values, multiplicity, value set, archetype reference
3. Terminology Bindings: archetype term definitions, references to standardized terminology resources like code systems, concepts, value set and value-set members

2.2.2 Create mappings between data dictionary elements and standard template elements

We then establish the mappings between corresponding data dictionary elements and the standard template elements. The mappings helped us observe how the data dictionary properties align to create a valid archetype and yet keep the semantics of the data model represented by a data dictionary intact. The informative section ‘AML-UML Transformation Reference’ of OMG AML specifications, which describes core set of mappings between ADL and AML, guided the creation of these mappings. As illustrated in Figure 2, a data dictionary is defined by and composed of its study variables. This structure resembles with the CIMI Core Reference archetype Cluster (as shown in Figure 1), which could be constrained to create almost any hierarchical collection. A data dictionary maps to an archetype, which is defined by constraining the reference archetype Cluster. Each
variable definition is mapped to a leaf-level constraint node (modeled by further constraining the Cluster.Element). All these archetypes are created to reside in an Archetype Library, which maps from the domain or study of the data dictionary.

2.2.3 Implement using D2Refine platform and tooling

We previously implemented a set of extensions to OpenRefine known as D2Refine. D2Refine includes features that can represent the Standard Template developed above. We plan to add a new extension to the D2Refine platform that allows us to serialize data dictionaries to and from AML archetypes, and a second to determine whether a given data dictionary meets the requirements. D2Refine provides the extensible reconciliation services needed to standardize, validate and transform data dictionary variable values.

2.2.4 Evaluate with a case study using a sample data dictionary

To test the utility of these mappings, we performed a case study and manually converted a sample data dictionary into AML archetypes. Figure 3 shows a customized dbGaP data dictionary with three variables. These three variables represent the most common types of constraints we observed in a data dictionary. The example shown here contains an identifier, a value set definition and an interval combined with a coded value.

3 Results

3.1 The Mappings

The results of mappings between data dictionary elements and standard template elements are shown in Tables 1, 2, and 3. The second column of each table lists the target properties of the Standard Template, which were derived from various UML stereotypes in three AML Specification profiles. The prefixes RMP, CP and TP are used to show the namespace of the AML UML profiles - the Reference Model Profile, Constraint Profile and Terminology Profile respectively. The data dictionary variables and their attributes are described using either pseudo-XPATH syntax or by their column names.

**Table 1. Mapping for the Archetype metadata**

<table>
<thead>
<tr>
<th>Mapping description for model metadata</th>
<th>Standard Template Elements (Follows OMG AML Specifications; in UML)</th>
<th>dbGAP (XML Schema)</th>
<th>PheKB (Spreadsheet)</th>
<th>TCGA (XML Schema)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The model library</td>
<td>CP.ArchetypLibrary.name</td>
<td>Study name</td>
<td>File name</td>
<td>Schema name</td>
</tr>
<tr>
<td>A model in the Library</td>
<td>CP.Archetype.name</td>
<td>data_table @study_id</td>
<td>TABLENAME</td>
<td>Schema name</td>
</tr>
<tr>
<td>Model definition</td>
<td>CP.ComplexObjectConstraint.name</td>
<td>data_table @study_id</td>
<td>TABLENAME</td>
<td>Schema name</td>
</tr>
<tr>
<td>Constrained RM class</td>
<td>CP.Constrains</td>
<td>Cluster is constrained</td>
<td>Cluster is constrained</td>
<td>xs:restriction</td>
</tr>
</tbody>
</table>

**Figure 3. A Sample dbGaP data dictionary**
Table 1 lists the first set of Standard Template elements for Archetype Metadata: Archetype library – a collection to which an archetype belongs, Archetype’s defining top-level constraint constraining RM Class or a parent archetype. The sub-constraints constraining RM Class properties follow the same pattern.

Table 2 shows the mappings satisfying requirements for data value types and value constraints like intervals and permissible values of a value set. In additional to the primitive types, the CIMI Core Reference Model defines all kinds of (concrete) Data Value Types like COUNT, QUANTITY, CODED_TEXT and others. All these data types are descendents of an abstract data value type DATA_VALUE. The target archetype constraint constrains an appropriate DATA_VALUE type, which depends on type of the values associated with a variable in the data dictionary. The ‘ArchetypeRoot’ constraint of an archetype is to create a reference to another archetype (composition).

### Table 2. Mappings for data types and value constraints

<table>
<thead>
<tr>
<th>Mapping description for constrained values and their data types</th>
<th>Standard Template Elements</th>
<th>dbGaP (XML Schema)</th>
<th>PheKB (Spreadsheet)</th>
<th>TCGA (XML Schema)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data types</td>
<td>RMP.MappedDataType data type or CIMI Reference Model data value type DATA_VALUE</td>
<td>variable/ type</td>
<td>TYPE</td>
<td>XML Primitive types, xs:element @ref</td>
</tr>
<tr>
<td>Values</td>
<td>Redefines or subsets the value of the constrained property (e.g. redefines Cluster.Element.value)</td>
<td>variable/ value</td>
<td>FORMATTED VALUE</td>
<td>xs:element @value</td>
</tr>
<tr>
<td>Encoded values</td>
<td>Constrains CIMI Reference Model data value type CODED_TEXT.code</td>
<td>variable/ value@code</td>
<td>RAW VALUE</td>
<td>xs:enumeration</td>
</tr>
<tr>
<td>Archetype Reference</td>
<td>CP.ArchetypeRoot</td>
<td>N/A</td>
<td>N/A</td>
<td>xs:element @ref</td>
</tr>
<tr>
<td>Interval Values</td>
<td>UML:Interval constraint</td>
<td>variable/ logical-min, logical-max</td>
<td>MIN, MAX, xs:element @minOccurs, xs:element @maxOccurs</td>
<td></td>
</tr>
<tr>
<td>Temporal Interval Values</td>
<td>UML:Duration constraint</td>
<td>variable/ logical-min, logical-max</td>
<td>REQUIRED, REPEATED MEASURE, xs:attribute @use</td>
<td></td>
</tr>
<tr>
<td>Multiplicity/ Occurrences</td>
<td>UML:MultiplicityElement</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3 lists the mappings for an archetype’s terminology bindings for the terms and codes identified from data dictionary variables. These mappings translate into references to internal and external vocabulary resources (e.g. NCI Thesaurus). The stereotypes of the Terminology Profile combine the features of ISO 11179-3 model and Common Terminology Services 2 (CTS2) specifications. The ISO 11179-3 model guides us for identification, designation, definition and value/meaning binding aspects. The CTS2 specification provides model for Concept, Code System, Code System Version, Value Set and Value Set Definition references.

**Table 3. Mappings for terminology bindings and value set references**

<table>
<thead>
<tr>
<th>Mapping description for terminology bindings</th>
<th>Standard Template Elements (Follows OMG AML Specifications; in UML)</th>
<th>dbGaP (XML Schema)</th>
<th>PheKB (Spreadsheet)</th>
<th>TCGA (XML Schema)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term Identification</td>
<td><strong>TP.ArcehtypeTerm.id</strong></td>
<td>variable@code</td>
<td>RAW VALUE</td>
<td>xs:enumeration@value</td>
</tr>
<tr>
<td>Term Text</td>
<td><strong>TP.IdEntry.text</strong></td>
<td>variable/value</td>
<td>FORMATTED VALUE</td>
<td>xs:elementt@name</td>
</tr>
<tr>
<td>Value Set Name</td>
<td><strong>TP.IdEntry.text</strong></td>
<td>variable/name</td>
<td>Worksheet</td>
<td>xs:elementt@name</td>
</tr>
<tr>
<td>Value Set Member Name</td>
<td><strong>TP.ArcehtypeTerm.value_set_members</strong></td>
<td>variable/name</td>
<td>Separate value-set Worksheet</td>
<td>xs:enumeration@value</td>
</tr>
<tr>
<td>Value Set Member Code</td>
<td><strong>TP.ArcehtypeTerm.id</strong></td>
<td>variable@code</td>
<td>RAW VALUE</td>
<td>xs:enumeration@value</td>
</tr>
<tr>
<td>Code System or Value Set Reference</td>
<td><strong>TP.CodeSystemReference, TP.CodeSystemVersionReference, TP.ValueSetReference, TP.ValueSetDefinitionReference</strong></td>
<td>NCI Thesaurus codes found, but other code systems could be used</td>
<td>DOCFILE + SOURCE + SOURCEID</td>
<td>caDSR</td>
</tr>
<tr>
<td>Concept or Value Set member Reference</td>
<td><strong>TP.ConceptReference, TP.PermissibleValue</strong></td>
<td>variable@code</td>
<td>RAW VALUE</td>
<td>attribute ‘cde’</td>
</tr>
</tbody>
</table>

These mappings were reviewed and verified by a panel of subject matter experts and the utility of the mappings was demonstrated by a successful transformation of a testing data dictionary into AML archetypes. Figure 4 shows a sample dbGaP data dictionary implemented as an AML archetype. As per mappings listed in the tables earlier, we used the AML stereotypes of ‘ArcheypeLibrary’, ‘Arcehtype’, ‘ComplexObjectConstraint’, ‘EnumeratedValueDomain’ for modeling Archetype Library, Archetype, constraints and terms (and terminology bindings) respectively. The archetype named ‘pht003255’ constrains the Cluster reference archetype. It is composed of three sub-constraints: ‘SUBJID’, ‘SEX’ and ‘AGE_FIRST’ – that are directly mapped from the data dictionary variable names. The target type of these constraints is mapped to the AML Primitive and Data Value
Types. Figure 4 shows the size and complexity of a resulting UML model for three simple constraint definitions (an online version available at ‡).

Figure 4. The sample dbGAP data dictionary implemented as an AML archetype

Please note that data dictionary variable identifiers are replaced with identifiers that follow the AOM identification scheme for clarity. The AOM specification describes identifier prefixes as id (normal identifier), ac (value-set identifier) and at (value-set member). The numerical number after these prefixes helps identify specialization level quickly.

3.2 Implementation status of the D2Refine transformation tools

The subsequent tasks, collectively known as D2Refine workbench 30, are already in-progress, to:

- Prepare OpenRefine platform for configuration and validation of the Standard Template elements
- Extend the import/export plugins of OpenRefine to load, transform and persist data dictionaries as AML archetypes

‡ https://raw.githubusercontent.com/caCDE-QAD2Refine/master/PrototypeProject/model_images/dbGapDDAMLModel/ph003255_Archetype.png
• We have already implemented a prototype transform that transforms dbGaP data dictionary to OpenEHR’s ADL format. Similar extension is required to implement the transformation to AML archetypes.

• Configure and extend OpenRefine’s reconciliation services with internal vocabulary terms and external terminological resources like NCI T, SNOMED CT, LOINC, caDSR and FHIR Profiles and their implementation.

• We have an Eclipse Modeling Framework (EMF) implementation of the CIMI Core Reference Model, which we have already been using.

Recently, we have successfully augmented OpenRefine’s reconciliation capabilities using any CTS2 compliant repository, which enables users to create terminology bindings of data dictionaries. We believe that we are on track to implement these features of D2Refine Workbench (Figure 5) for real-world use cases in near future. With the workbench support, we can practically and effectively map all metadata of any data dictionary with the Standard Template elements, CIMI Core Reference Model and the CIMI Core Reference archetypes.

4 Discussion

Data dictionaries provide simplest form to capture and list metadata entries in a tabular format. The tabular format is usually derived from a spreadsheet, a structured XML document or a set of delimited values arranged as row and columns. This simple form, though providing fastest way to organize and understand the datasets, can be really diverse and rarely portable outside of their environment. The encouraging aspect of data dictionary is that they are all very similar in the way of capturing the details of the variables. A data dictionary lists variable definitions with their names, data types, value ranges and binding to local and external resources, which is analogous to how an archetype is composed. A data dictionary does not have a reference model, which is needed by an archetype to define constraints. We found that the reference model component of an archetype definition can easily be substituted either by organizing structures of data dictionaries into a reference model classes or utilizing an existing reference model that is sufficient. In addition, the data dictionary formats used in this study are identified from three large projects: dbGap, TCGA and PheKB. dbGap and TCGA are two NIH pilot data commons; and PheKB is a catalog of electronic phenotype algorithms and associated data dictionaries, and largely used in the eMERGE Research Network. We believe that the selected data dictionary formats represent well the current practice in the clinical research communities.

The simplified view of CIMI Reference Archetype - Cluster, shows a recursive structure definition by constraining CIMI Reference Model (RM) classes ITEM_GROUP and ELEMENT for grouping or representing leaf level entities respectively. Since data dictionaries do not have a reference model specified, we use CIMI Core Reference Archetypes and constrain them to model data dictionary constraints. In case of the TCGA data dictionary where a number of commonly used structure definitions were introduced, these definitions were augmented to the collection of CIMI Core RM types and the reference model extended into a TCGA Reference Model. The creation of TCGA RM made it possible to correctly map the TCGA data dictionary constraints to the standard template requirements.

The implementation of standard template is not only able to preserve the existing tabular, spreadsheet-like interface to create and manage data dictionaries, but also facilitates additional transforms to better utilize them in different environments. The standard template directly maps from AML profiles and hence guides the modelers to make sure that their models have the necessary model elements and could be persisted as archetypes. The ‘archetype modeling’ is the ‘constraint-based modeling’ where constraints are about the reference model classes. The Archetypes stand separate from the reference model that does not change in any way, which introduce flexibility and improve interoperability.

Figure 5. The D2Refine Workbench
Besides AML, the Archetype Definition Language (ADL)\(^1\) is the other supported formalisms approved by the HL7 CIMI for describing clinical models. A model in ADL format is described in Object Data Instance Notation (ODIN) text whereas an AML Model is represented in UML. We are also actively developing tools that enable the transformation of data dictionaries into the ADL representation. The transformed data dictionaries into the ADL format can leverage the tools and environment developed for ADL models and that can help harmonize them with thousands of existing ADL archetypes. On the other hand, the transformation of data dictionaries into CIMI archetype’s AML format (a non-proprietary UML model) is an important gateway to the Model-Driven Architecture (MDA)\(^3\). The MDA workflow provides a way to expedite development of flexible and robust dataset validation tools and user applications. The AML specifications are guided by ADL Object Model (AOM) and include transformation mappings to seamlessly move between the ADL and AML. These mappings between ADL and AML are being implemented by the community and will be available in near future.

The standard template in D2Refine preserves the spreadsheet-like tabular interface that allows users to review data dictionaries; at the same time it allows model metadata to bind to standard terminologies and metadata repositories with its extensible reconciliation services. D2Refine also provides built-in framework for implementing bi-directional transforms to store and manage data dictionaries using AML formalism.

The task of creating CIMI AML archetypes requires implementing the AML Profiles and using UML implementation to create AML’s UML artifacts. The open-source EMF\(^2\) and Eclipse UML2 Libraries are essential to this task. A library of convenient programming interfaces encapsulating the EMF and UML2 implementations is being developed. Ability to access these interfaces from OpenRefine extensions should be sufficient to realize and persist AML archetypes.

The manual process of creating the AML archetypes from data dictionaries is impractical, tedious and error-prone. As described earlier, we will need various items to programatically transform data dictionaries. The open-source EMF implementation of the AML Profiles is crucial to be able to create AML objects. We have worked on investigating the Model Driven Health Tools (MDHT)\(^4\), an Eclipse open-source project, to generate AML profiles implementation (using OMG AML specifications) as a library. We plan to use Eclipse’s UML2 libraries to create UML 2.5 objects.

5 Conclusions

In this study, we successfully created reliable mappings between three types of data dictionaries and the standard template elements informed by the AML specification. We have demonstrated that the mappings are very helpful in enabling the representation of a data dictionary in a CIMI archetype. We are actively implementing the mappings in a D2Refine platform\(^5\) to enable the transformation of data dictionaries into CIMI archetypes. The outcome of our work will enable the standard representation of heterogeneous clinical study data dictionaries, thereby facilitating effective metadata harmonization and downstream data integration, ultimately advancing clinical research studies.

Acknowledgement

This work was supported in part by funding from R01 GM105688, R01 GM103859 and a NCI U01 Project – caCDE-QA (U01 CA180940).

References

Identifying Patients at Risk of High Healthcare Utilization

Lincoln Sheets,¹,² MD, Lori Popejoy,³ PhD, APRN, GCNS-BC, Mohammed Khalilia,¹,⁴ PhD, Greg Petroski¹, PhD, Jerry C. Parker¹, PhD

1. University of Missouri School of Medicine
2. University of Missouri Informatics Institute
3. University of Missouri School of Nursing
4. University of Missouri College of Engineering

This publication was made possible by Grant Number 1C1CMS331001-01-00 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies. The funding agreement ensured the authors’ independence in designing the study, interpreting the data, writing, and publishing the report. The research presented was conducted by the awardee. Findings may or may not be consistent with or confirmed by the findings of the independent evaluation contractor.

This project was funded by the Center for Medicare and Medicaid Services (CMS) to expand the scope of services to a population of CMS beneficiaries, so the Health Sciences Institutional Review Board deemed the project to be a quality improvement initiative that did not require a formal patient consent process since the explicit purpose of data use was to improve patient care; the IRB number is 1212477.QI. The authors have no conflict of interest to declare.

ABSTRACT

Objective. To develop a systematic and reproducible way to identify patients at increased risk for higher healthcare costs. Methods. Medical records were analyzed for 9,581 adults who were primary care patients in the University of Missouri Health System and who were enrolled in Medicare or Medicaid. Patients were categorized into one of four risk tiers as of October 1, 2013, and the four tiers were compared on demographic characteristics, number of healthcare episodes, and healthcare charges in the year before and the year after cohort formation. Results. The mean number of healthcare episodes and the sum of healthcare charges in the year following cohort formation were higher for patients in the higher-risk tiers. Conclusions. Retrospective information that is easily extracted from medical records can be used to create risk tiers that provide highly useful information about the prospective risk of healthcare utilization and costs.

INTRODUCTION

The Chronic Care Model¹ proposed improving the effectiveness of interactions between patients and providers as a way of promoting the “triple aim” of healthcare:² (a) better health, (b) better care, and (c) lower costs.³ These positive outcomes may be anticipated on the basis of improved interactions between informed, activated patients and prepared, proactive providers ¹. Evidence also has shown that patients with care coordinators have fewer emergency department and urgent care episodes,⁴ hospital admissions,⁵ and readmissions.⁶
Care coordination is “… the deliberate organization of patient care activities between two or more participants (including the patient) involved in the patient’s care to facilitate the appropriate delivery of health service…” Care coordination is increasingly being used across the healthcare system to improve patient outcomes for populations of patients and is a core element of both the triple aim and the Chronic Care Model. In order for care coordinators to manage populations of patients, they must be able to identify the patients who are most in need of their services; one approach to this challenge is risk stratification. Indeed, risk stratification is crucial for effective population health management because it provides care coordinators the opportunity to focus their work on those patients who will benefit the most. By bridging the implementation gaps in the Chronic Care Model, well-designed risk stratification supports the transition from the traditional “reactive” model of medical care to one of maintaining health and avoiding preventable conditions. Risk stratification is a potentially powerful tool for predicting population health outcomes, and previous healthcare utilization has been shown to be a useful predictor of future healthcare needs.

Although the literature contains many studies and reviews of condition-specific risk stratification, most predictive indices have less than 90 percent accuracy and relatively few attempts have been made to stratify entire primary care populations. Notable examples include the Scottish SPARRA risk index for readmission of patients who had been recently discharged from the hospital and the Welsh Prism risk index for emergency admissions. In the United States, the Michigan Primary Care Transformation Project defined four functional tiers upon which intensity of services was based, rather than relying on a predictive index. The tiers are as follows: (a) Navigating the Medical Neighborhood, (b) Transition Care, (c) Care Management, and (d) Complex Care Management. The Senior Segmentation Algorithm, developed by the Kaiser Permanente healthcare system, also used four tiers that were more descriptive than functional; those were (a) Without Chronic Conditions, (b) With One or More Chronic Conditions, (c) With Advanced Illness or End-organ Failure, and (d) With Extreme Frailty or Nearing the End of Life. The Senior Segmentation Algorithm, while objective and reproducible, has been validated only in patients aged 65 years and older. Other models that have been applied to community-dwelling populations with varying degrees of success are Dr. Chad Boult’s work on risk scores to identify high risk patients from Medicare data, the Charlson and Elixhauser comorbidity indices, and the work by Dr. Hal Luft on the yearly stability of Medicare utilization and costs.

In this study, the objective was to develop a systematic, reproducible way to identify subgroups in a managed population at risk for higher utilization of healthcare resources as measured by healthcare episodes and charges.

**METHODS**

**Population**

In February of 2013, the LIGHT2 project (Leveraging Information Technology to Guide Hi-Tech and Hi-Touch Care, pronounced “light squared”) began enrolling adults who were primary care patients in the University of Missouri Health System and who were already enrolled in Medicare and/or Medicaid. LIGHT2 was a Health Care Innovation Award from CMS (Centers for Medicare and Medicaid Services) to achieve the Triple Aim of better health, better care, and lower costs by using advanced information technology and care coordination. In order to define a stable cohort,
this study limited data analysis to 9,581 patients who were enrolled before the first three months of care coordination was completed on July 1, 2013, and who were still enrolled when patients were first evaluated for risk stratification on October 1, 2013.

Primary Care Setting

Adult primary care in the University of Missouri Health System (UMHS) is provided by approximately 133 primary care physicians who practice in nine regional clinics; the Department of Family and Community Medicine operates five local outpatient clinics and two in nearby communities, and the General Internal Medicine section of the Department of Medicine operates two local clinics. This community-based, primary care focus is supported by an extensive UMHS tertiary-care system of six hospitals and more than 50 clinics, staffed by approximately 550 university physicians.

Data Source

All data on diagnoses, outpatient visits, and hospital episodes and charges were retrieved from the University of Missouri Health System electronic medical record as maintained by clinicians and staff between 2012 and 2014.

Rationale for Categorical Approach

For the purpose of differentiating levels of care coordination intensity, a categorical approach has some advantages over an index. Each category can receive a customized level of care coordination in a manner similar to the “functional tiers” of the Michigan Primary Care Transformation Project. One potential disadvantage to some previous approaches is that they sometimes rely on socioeconomic and other data for which characteristics can vary widely across settings. In order to predicate the risk tiers on more universally reproducible criteria, socioeconomic factors were not included in the development of risk tiers for this study using the general criteria of diagnoses and utilization. Specifically, the 27 chronic conditions that are included in the Chronic Conditions Data Warehouse provide a standard set of diagnoses that are relevant to population health management of Medicare/Medicaid patients in a primary care setting (see Figure 1). In addition, healthcare utilization was extracted for (a) outpatient visits, which may be encouraged or even increased as part of care coordination, and (b) the more costly hospital-based episodes, which include emergency department episodes, observation stays, and hospital admissions.

Number and Definitions of Tiers

Tier Definitions. The definitions of the four LIGHT2 risk tiers are shown in Table 1. These tiers were conceptualized to be generally similar to the “functional tiers” developed by the Michigan Primary Care Transformation Project and the “care groups” used in the Senior Segmentation Algorithm. For the purposes of risk stratification, the LIGHT2 tiers were defined as (1) No Chronic Conditions, (2) Chronic Conditions, Stable, (3) Chronic Conditions, Unstable, and (4) Chronic Conditions, Complex. These broad categories of chronic-disease acuity describe all members of a managed adult population with some specificity, but the categories also are general enough to be applied to a variety of environments and conditions.

Tier Calculations. The breakpoints between tiers were determined deductively in an effort to
reflect the intent of the functional tier definitions. The lowest or “No Chronic Conditions” category (Tier 1) was reserved for enrollees who had none of the 27 diagnoses included in the CMS Chronic Conditions Data Warehouse. For those enrollees with one or more chronic conditions, placement in either Tier 2, 3, or 4 was based on the frequency of outpatient clinic visits and hospital episodes (including emergency department episodes, observations stays, and hospital admissions) during the 12 months prior to analysis. Patients with one or more of the 27 CMS chronic conditions were stratified as “Chronic Conditions, Stable” (Tier 2) if they had four or fewer outpatient visits and no hospital episodes associated with their chronic conditions in the preceding 12 months. Based on the clinical experience of two authors (LS, LP), patients were stratified as “Chronic Conditions, Unstable” (Tier 3) if they had 5 to 12 outpatient visits or one hospital episode related to any chronic conditions during the preceding 12 months, and were stratified as “Chronic Conditions, Complex” (Tier 4) if they exceeded 12 related outpatient visits or had two or more related hospital episodes during the preceding 12 months.

Data Collection

Cohort Formation. In order to test the ability of these queries to predict risk for higher utilization of healthcare resources, the population was divided into tiers based on their clinical history as of a given date (October 1, 2013) and was treated thereafter as four fixed cohorts. For each cohort, the healthcare episodes and charges were calculated retrospectively for the 12 months before cohort formation (October 1, 2012, to September 30, 2013) to provide a baseline, and then prospectively for the subsequent 12 months (October 1, 2013, to September 30, 2014) to examine the utility of the risk stratification methodology.

Cohort Attrition. Attrition (e.g. death, relocation) during the 12 months after cohort formation reduced the size of the cohort by 208 patients, or 2.2%. Because excluding these patients was likely to bias the data collection toward healthier patients, their outcomes were included for those months in which they were alive and enrolled in the LIGHT2 program. A patient enrolled for one month after cohort formation, therefore, would contribute one-twelfth as much to the mean episodes and charges for his or her tier subgroup as another patient who was enrolled for the entire 12 months.

Inflation Adjustment. In order to compare baseline and prospective charges, all charges were adjusted to 2014 dollars. Since healthcare charges were adjusted upward by 3% on April 1 of each year during the study period, charges billed from April 1, 2013, to March 31, 2014, were adjusted by multiplying times 1.03. Charges billed from October 1, 2012, to March 31, 2013, were multiplied by (1.03 x 1.03) or 1.061.

Statistical Methods

Statistical comparisons of outcomes between tiers used the Kruskal-Wallis Test, a nonparametric analogue of the standard one-way ANOVA. Statistically significant (p < 0.05) overall tests were followed by pairwise rank-based comparisons; all comparisons were two-sided.

RESULTS

Demographic Description of the Cohort
Of 9,581 patients in the cohort, 63% (n = 6,014) were in Tier 2 on October 1, 2013, while Tiers 1 and 3 comprised 16% each (n = 1,554 for Tier 1; n = 1,555 for Tier 3), with the remaining 5% (n = 458) in Tier 4. More than three-fourths of the Tier 1 enrollees were younger than 65, but fewer than half of the enrollees in the other tiers were under 65. Overall, 4,185 of the 9,581 patients (44%) were younger than 65 years old. Approximately three-fifths of the enrollees in every tier were female. Figure 1 shows the prevalence of the Chronic Conditions Data Warehouse diagnoses in the LIGHT2 population.

Baseline Analyses of Hospital Episodes by Tier

During the 12 months before cohort formation, the mean number of emergency department episodes was 0.20, 0.18, 0.52, and 1.45 in each tier respectively; observation stays averaged 0.10, 0.09, 0.41, and 1.16 respectively; and there were 0.09, 0.09, 0.47, and 1.86 mean inpatient admissions in each respective tier. The number of episodes was significantly different overall among the tiers (p < 0.001) and progressively higher in successive tiers in pairwise comparisons (p < 0.001 for all comparisons), except that there were no significant differences between numbers of episodes in Tiers 1 and 2 (p = 0.304 for emergency, p = 0.967 for observation, and p = 0.739 for inpatient). Because the data were skewed, the median number of episodes of all types was zero for Tiers 1, 2, and 3. For Tier 4, the median number of emergency episodes was zero, and the median number of observation and inpatient episodes was 1. The median values of zero indicate that over half of the lower tier patients had no hospital episodes during the baseline year.

Baseline Analyses of Healthcare Charges by Tier

During the 12 months before cohort formation, mean healthcare charges in the four tiers (respectively $6,208, $10,889, $43,059, and $115,228) were significantly different overall and in pairwise comparisons (p < 0.001 overall and for all comparisons). Median charges in each tier (respectively $658, $2,641, $19,707, and $66,182) were likewise significantly higher in higher tiers (p < 0.001 overall and for all pairwise comparisons). Tiers 3 and 4, comprising only 21% of the total population, accounted for 61% of total healthcare charges. Because lower tiers were defined by fewer or no healthcare episodes, the finding that enrollees in these tiers had fewer mean hospital episodes and lower charges was expected. However, these numbers show the magnitude of the differences between the subgroups in each tier at baseline, and they provide the basis for tier categorization on the date of cohort formation (October 1, 2013).

Prospective Analyses of Hospital Episodes by Tier

Figure 2 shows the mean number of hospital episodes by type, within each tier, during the 12 months following cohort formation. For all episode types, the overall differences remained significant (p < 0.001). Tier 4 enrollees had significantly more episodes on average than Tier 3 patients; they, in turn, had significantly more episodes than Tier 2 or Tier 1 patients (p < 0.001 for all comparisons). There were no significant differences between Tier 1 and Tier 2 emergency episodes (p = 0.279), but the numbers of observation and inpatient episodes were significantly different between these two tiers (p < 0.001 for each type). For all tiers, the median number of episodes of all types was zero.

Prospective Analyses of Healthcare Charges by Tier

1133
Figure 3 shows the mean and total healthcare charges by tier during the 12 months after cohort formation. Tiers 3 and 4, comprising 21% of the total population, accounted for 43% of total healthcare charges. Overall differences were significant (p < 0.001). Tier 4 enrollees had significantly higher charges on average than Tier 3 patients, which had significantly higher charges than Tier 2 patients; and these in turn where significantly higher than for Tier 1 patients (p < 0.001 for all comparisons). Median charges in each tier (respectively $0, $2,343, $8,662, and $20,412) were likewise significantly different overall (p < 0.001) and higher in the higher tiers (p < 0.001 for all comparisons).

DISCUSSION

The primary conclusion is that the LIGHT2 risk stratification methodology successfully met the objective of identifying which subgroups of Medicare/Medicaid patients were at risk of utilizing more healthcare resources. Specifically, information about healthcare utilization and charges over the previous twelve months provided highly useful information about what is likely to occur going forward. This finding confirms the expectation that past utilization is a good predictor of future utilization. When viewed through a population health lens, the extraction of readily available retrospective data can provide care coordinators with useful information that allows them to focus their efforts on those patients who care needs are most expensive, and who may require more intensive management.

A second observation is that the risk tiers created in this study were defined on the basis of diagnoses that are included in the Chronic Conditions Data Warehouse and on utilization categories that are in general use in healthcare systems. Because these are easily reproducible, the LIGHT2 risk stratification framework is potentially applicable to other healthcare systems.

A third observation is that there were no differences between Tier 1 patients and Tier 2 patients in the mean number of hospital episodes during the retrospective (i.e., baseline) period. Similarly, in the prospective period, there was no difference in mean number of emergency department episodes between Tiers 1 and 2. Accordingly, the cost and utilization patterns of Tier 2 patients are, in some ways, not notably different from Tier 1. Therefore, from a population health standpoint, time and resources devoted to keeping Tier 2 patients from evolving into the much more costly Tier 3 category would appear to be a useful care coordination strategy.

Finally, Tiers 3 and 4, comprising only 21% of the total population, accounted for 61% of total baseline healthcare charges and 43% of total prospective healthcare charges. The decrease in this difference from the baseline to the prospective measure reflects that prior healthcare episodes and charges are not a perfect, but nevertheless, highly useful predictor of prospective utilization. Specifically, the consumption of a large proportion of healthcare resources by a relatively small minority of the population (Tiers 3 and 4), even in the prospective year, underscores the importance of the early identification of those patients (or subgroups) who are most “at risk” for high healthcare utilization.

Limitations

Measurement of baseline and prospective utilization included only healthcare episodes that occurred within the University of Missouri Health System. Accordingly, these results likely underreport utilization by excluding healthcare episodes at other local or out-of-town facilities; inclusion
of this additional data could possibly improve the accuracy of risk stratification. In addition, healthcare charges were used as a proxy for healthcare costs, although claims data would be a more accurate source of cost information. Lastly, the generalizability of the LIGHT2 risk stratification methodology to other settings remains an empirical question. The study sample included only adults who received Medicare or Medicaid services in central Missouri, which may limit the generalizability of these findings to the wider US population.

**Future Research**

This preliminary work would be strengthened by validation against other populations that may have different prevalence of chronic conditions, age-adjusted death rates, or rates of relocation as well as different distributions of race/ethnicity, educational attainment, economic status, and literacy. Because the data used in the LIGHT2 risk stratification methodology is available in any healthcare system, replication of this four-tier methodology in other settings would be instructive. Additional studies that incorporate claims data also could be conducted to examine whether the accuracy of risk stratification could be improved. While this study demonstrates the predictive power of a tiered model, continuous risk scoring may be even more powerful, and development of such a model would also be helpful.

**ACKNOWLEDGMENTS**

This publication was made possible by Grant Number 1C1CMS331001-01-00 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies.

The authors gratefully acknowledge the editorial assistance of Cynthia Haydon, University of Missouri.

**REFERENCES**

7. McDonald G, Vickers MH, Mohan S, Wilkes L, Jackson D. Workplace Conversations:
TABLES

Table 1: Operationalized Queries

<table>
<thead>
<tr>
<th>Tier</th>
<th>Definition (based on healthcare episodes in the past 12 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: No Chronic Conditions</td>
<td>Zero chronic conditions listed in the Chronic Conditions Data Warehouse</td>
</tr>
<tr>
<td>2: Chronic Conditions, Stable</td>
<td>One or more chronic conditions AND (hospital episodes = 0 AND outpatient visits &lt; 5)</td>
</tr>
<tr>
<td>3: Chronic Conditions, Unstable</td>
<td>One or more chronic conditions AND (hospital episodes = 1 OR outpatient visits from 5 to 12)</td>
</tr>
<tr>
<td>4: Chronic Conditions, Complex</td>
<td>One or more chronic conditions AND (hospital episodes &gt; 1 OR outpatient visits &gt; 12)</td>
</tr>
</tbody>
</table>

FIGURES

Figure 1: Prevalence of 27 Defined Chronic Conditions

![Figure 1: Prevalence of 27 Defined Chronic Conditions](image-url)
Figure 2: Prospective Hospital Episodes by Tier

![Bar chart showing the mean number of hospital episodes per patient by tier.]

Tier 1: Emergency 0.23, Observation 0.07, Inpatient 0.23
Tier 2: Emergency 0.13, Observation 0.15, Inpatient 0.23
Tier 3: Emergency 0.39, Observation 0.32, Inpatient 0.25
Tier 4: Emergency 1.01, Observation 0.61, Inpatient 1.08

Figure 3: Prospective Healthcare Charges by Tier

![Bar chart showing total charges and mean charges per member by tier.]

Tier 1: $4,931 (5%)
Tier 2: $14,234 (53%)
Tier 3: $27,712 (27%)
Tier 4: $55,735 (16%)

Mean Charges:
- Tier 1: $7.7M
- Tier 2: $43.1M
- Tier 3: $25.5M
- Tier 4: $60,000
Knowledge as a Service at the Point of Care

Jane L. Shellum, MHA*; Robert R. Freimuth, PhD*; Steve G. Peters, MD; Rick A. Nishimura, MD; Rajeev Chaudhry, MBBS, MPH, PCIM; Steve J. Demuth; Amy L. Knopp, MSHS; Timothy A. Miksch, MBA; Dawn S. Milliner, MD
Office of Information and Knowledge Management, Mayo Clinic, Rochester, MN

Abstract
An electronic health record (EHR) can assist the delivery of high-quality patient care, in part by providing the capability for a broad range of clinical decision support, including contextual references (e.g., Infobuttons), alerts and reminders, order sets, and dashboards. All of these decision support tools are based on clinical knowledge; unfortunately, the mechanisms for managing rules, order sets, Infobuttons, and dashboards are often unrelated, making it difficult to coordinate the application of clinical knowledge to various components of the clinical workflow. Additional complexity is encountered when updating enterprise-wide knowledge bases and delivering the content through multiple modalities to different consumers. We present the experience of Mayo Clinic as a case study to examine the requirements and implementation challenges related to knowledge management across a large, multi-site medical center. The lessons learned through the development of our knowledge management and delivery platform will help inform the future development of interoperable knowledge resources.

Introduction
A core business of Mayo Clinic is the creation and application of integrated knowledge to deliver safe and high value care. As the first integrated not-for-profit medical group practice in the world\(^1\) Mayo Clinic pioneered a team-based multidisciplinary approach to medicine. That approach evolved into a system where all Mayo physicians relied routinely on direct interaction with colleagues from across the practice as the primary source for vetted, trusted clinical knowledge. As EHR-based automation and payment constraints put pressure on the formal structures that enabled this cultural approach, we sought to create a knowledge management program that captured Mayo’s collective, vetted practice knowledge as a clinical knowledge resource for our medical staff. The size of Mayo's clinical staff, which includes 4,200 physicians and scientists, 2,400 residents and fellows, and nearly 53,000 allied health personnel, as well as the geographic distribution of the practice (campuses in Minnesota, Arizona, and Florida as well as more than 70 communities within the Mayo Clinic Health System), makes the collection, curation, delivery, and maintenance of knowledge highly challenging. To help address these challenges, the Office of Information and Knowledge Management (OIKM) was developed in 2012. The Office, led by the Chief Medical Informatics Officer, is made up of the physicians and administrators responsible for information management, practice analytics, applied clinical informatics, and knowledge management and delivery. OIKM is responsible for creating the infrastructure, collaborative tools and methodologies required to manage our knowledge assets, perpetuate best practices, and deliver the right information seamlessly to providers and patients when and where it is needed.

AskMayoExpert
Clinicians in many settings fail to find answers to clinical questions many times each day\(^2\). Lack of time is reported to be the biggest barrier to finding such answers\(^3\). Nearly half of the time, clinicians don’t pursue answers, most often citing the expectation that no useful information would be found\(^4\). In response to these challenges, the AskMayoExpert (AME) application was developed to provide answers to clinical questions at the point of care. AME is based on the concept of medical “gist” – concise, relevant, and clinically applicable answers to clinical questions\(^5\). The application has been widely adopted and is an integral part of the practice\(^6\). It comprises more than 13,000 structured, concise, and actionable “knowledge elements” covering 1,300 clinical conditions. AME is available for both desktop browsers and mobile devices to clinicians at Mayo Clinic, within the Mayo Clinic Health System, and at sites within the Mayo Clinic Care Network. A controlled crossover study found AME enabled clinicians to answer clinical questions with greater accuracy and confidence than other resources\(^7\).

Generating Consensus-Based, Vetted Clinical Knowledge
At the same time that AME was being built, Mayo Clinic launched a significant initiative aimed at practice standardization. “Specialty Councils” were assembled to develop consensus regarding best practices and to converge, integrate and propagate vetted Mayo Clinic knowledge. AME was identified as the repository for...
standardized knowledge and best practices developed by the Specialty Councils. A knowledge management team works with the Specialty Councils to create interactive algorithms called Care Process Models (CPMs) to represent these best practices. CPMs are designed to guide a clinician through stepwise care of a patient with a suspected or identified condition, and to provide concise, actionable recommendations. They are delivered as static or semi-interactive web pages that can be embedded in or linked to from within AME content. In 2015, 155 CPMs were accessed 82,649 times. On average, a CPM is retrieved 10.2 times per week; the most commonly accessed CPM in 2015 was for hyperlipidemia (Figure 1).

One of the greatest challenges in building AME was developing the process for creation and capture of clinical knowledge that would assure its credibility and acceptance across the practice. Knowledge Content Boards (KCBs), groups of four to six physicians focused on a particular specialty area, are responsible for capturing best practices identified by Specialty Councils and vetting the AME content. The KCB members are given dedicated time, and their work is recognized as a scholarly activity for consideration in academic appointments. Currently, 234 clinicians participate in 38 different KCBs, directed by 47 Specialty Councils.

One of the greatest challenges in building AME was developing the process for creation and capture of clinical knowledge that would assure its credibility and acceptance across the practice. Knowledge Content Boards (KCBs), groups of four to six physicians focused on a particular specialty area, are responsible for capturing best practices identified by Specialty Councils and vetting the AME content. The KCB members are given dedicated time, and their work is recognized as a scholarly activity for consideration in academic appointments. Currently, 234 clinicians participate in 38 different KCBs, directed by 47 Specialty Councils.

Figure 1. Care Process Model (CPM) utilization in 2015

Core Clinical and Structured Knowledge

The body of knowledge created and maintained in AME was recognized as an asset with value beyond a reference at the point of care, and it became the basis for a formal knowledge management program charged with the creation, maintenance, and governance of a core body of clinical knowledge which can be delivered in multiple forms to diverse audiences. To support this goal, we established the concept of foundational or “core” clinical knowledge, from which other knowledge assets may be derived. For example, the “core clinical content” in AME now serves as the basis for derivative products such as patient education materials and consumer health information. It also serves as the basis for a broad range of clinical decision support mechanisms, including Infobutton content, alerts and reminders, and order sets (Figure 2).

AME has thus evolved to serve three purposes: answering clinical questions, representing care processes, and serving as the basis for other knowledge assets. Each of these purposes generated unique requirements for both the content and how that content was stored. To be useful in answering clinical questions, the content must be clear, concise, actionable, findable, and trusted. To represent CPMs, the content had to be categorized into decisions and actions that could be related visually, while at the same time ensuring that the content answers focused clinical questions. Introducing the notion of core clinical content from which other assets could be derived necessitated even more consistent structuring, richer annotation, and explicit relationships between content items.
To accomplish this, an in-depth analysis of the core content was performed by the content consumers and content structure specialists. A very detailed structure was proposed; however, testing found that while the structure accommodated a broad range of content, fragmentation reduced the usefulness in answering clinical questions or representing the care process (Table 1). Furthermore, while subject matter experts and support staff were advised that the structure was only a categorization scheme rather than a template to be completed, there was a temptation to “fill in the blanks,” thereby reducing the focus on the intended clinical content.

First Approach to Structuring (three separate entries)

<table>
<thead>
<tr>
<th>Medical service name</th>
<th>Condition-specific medical service detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitral valve stenosis-Transthoracic echocardiography (TTE)</td>
<td>Transthoracic echocardiography (TTE) is indicated for the initial evaluation of all patients with suspected mitral valve stenosis.</td>
</tr>
<tr>
<td>Mitral valve stenosis-Transthoracic echocardiography (TTE)</td>
<td>In addition to confirming the diagnosis of mitral valve stenosis, transthoracic echocardiography (TTE) is used to characterize mitral valve stenosis and direct treatment of the condition. If TTE does not adequately define the severity of mitral valve stenosis or identify concomitant mitral regurgitation or cardiac thrombus, transesophageal echocardiography is indicated.</td>
</tr>
<tr>
<td>Mitral valve stenosis-Transesophageal echocardiography (TEE)</td>
<td>Transesophageal echocardiography (TEE) is indicated to direct treatment of mitral valve stenosis if transthoracic echocardiography does not adequately define the severity of stenosis or identify concomitant mitral regurgitation or cardiac thrombus, particularly when considering percutaneous balloon valvuloplasty.</td>
</tr>
</tbody>
</table>

Revised Approach to Structuring (single, context-preserving entry)

<table>
<thead>
<tr>
<th>Medical service name</th>
<th>Condition-specific medical service detail</th>
</tr>
</thead>
</table>
| Mitral valve stenosis-Imaging Tests | The presence of mitral valve stenosis is confirmed with transthoracic echocardiography, which allows:  
  - Evaluation of valve morphologic characteristics and severity of stenosis  
  - Evaluation of sequelae of left atrial enlargement and pulmonary hypertension  
  - Detection of other valve abnormalities  
  Selected patients may require transesophageal echocardiography to confirm the severity of mitral valve stenosis or to identify concomitant mitral regurgitation or cardiac thrombus (particularly if percutaneous balloon valvuloplasty is being considered). |

Table 1. Content Structuring – Comparison of Approaches
A subsequent external review (see below) recommended that the content be structured to the business purpose; that is, make the content as granular as the end user can tolerate, but no more so. As a result, the knowledge structuring framework was simplified to include broad categories. A "universal template" was constructed based on the principles behind Darwin Information Typing Architecture (DITA), including modularity and reuse, information typing, and separation of content from form. This template provides a common, consistent framework that makes it easier to model, ingest and reuse content. The universal templates follow a hierarchical structure of the form Topic-Concept-Section, where the Section is the smallest unit of content that can be delivered on its own. Each Topic, Concept, and Section is given a unique content identifier and can be maintained independently. To date, the universal template has been instantiated to create two "anchor topics": Condition and Procedure. The Concepts approved for use within each of those Topics are shown in Figure 3. To date, the KCMS contains 504,349 items of content (of which 117,349 items are related to Mayo Clinic employees) and the AME contains 1541 published topics and 12,229 published sections. In the 6 month period from September 2015 to February 2016, 55,795 AME users initiated 129,270 sessions and viewed 587,768 pages of content (averaging more than 3200 page views per day).

**Figure 3.** Instances of the universal template: Concepts for the Topics "Condition" and "Procedure"

### Derived Knowledge Assets

The "core clinical content" serves as a single "source of truth", from which other knowledge assets can be derived. Derived assets, which are produced manually, take two forms: Structured narrative content including patient education materials and consumer web content, and computable assets including rules, assessment tools, and order sets. The former are authored in KCMS. The latter are cataloged in the KCMS but instantiated in the EHR and other clinical systems. Representing these computable assets generates multiple components which are themselves knowledge assets, including value sets, calculations, and cohort definitions. (Figure 4). For example, the CHA2DS2-VASc score calculator, which is embedded within the CPM for anticoagulation in patients with atrial fibrillation, utilizes cohort definitions in addition to several clinical parameters. The catalog records for computable knowledge assets serve as a common, enterprise-wide definition that can be instantiated within local systems. For example, the "clinical decision support rule" knowledge assets are stored in a master rule catalog as structured algorithms in system-agnostic syntax. When those algorithms are implemented across our varied EHR systems as executable rules, metadata about those implementations are added to the catalog to provide a link between the centralized knowledge asset and the implemented rule.

The core content serves as a starting point for generating a new knowledge asset, but additional knowledge acquisition is required for each derived item since it is developed for a particular purpose. For example, a decision support rule requires the extraction of computable logic and defined clinical concepts from the narrative core content. An order set requires details about specific orderable items and unit routines. Patient educational materials require the addition of broader background content as well as directions for self-care. The core content is written primarily for a physician/advance practice nurse audience. A pilot is under way to add nursing-focused content. In general, the core content has sufficient depth such that additional content can be generated by clinical informaticians and other members of the care team (e.g. nurses, pharmacists, therapists) with minimal input from the subject matter experts.
While the early emphasis of the knowledge management program was on narrative or reference content specific to a condition or clinical situation, but not to a particular patient, we knew that in order to effectively apply knowledge to patient care it must be individualized to a particular patient. We also recognized that, even though electronic health records (EHRs) automate clinical data access and workflow – and thus are a tempting place to directly embed clinical knowledge – clinical workflow is fundamentally different from the medical knowledge and decision support assets that are the target of our knowledge management program. While knowledge buried in purely operational workflow is necessarily implicit, and thus difficult to manage and sustain, knowledge represented only in human-readable artifacts is not routinely actionable and is difficult to scale. Thus, an architecture in which clinical knowledge is separated from but integrated with the patient data and workflow delivery systems is critical to the ability to truly scale knowledge delivery.

A dedicated knowledge content management system (KCMS) serves as a primary source for clinical knowledge. The KCMS serves three functions: a) it provides the authoring environment for narrative content, b) it delivers content and search functionality on demand to users, and c) it functions as a catalog for other types of knowledge assets. Other systems that make use of the clinical knowledge in the KCMS can be considered as either knowledge staging or implementation systems. A knowledge staging system delivers knowledge content such as text, formatted text, images, or video to knowledge consumers. Staging systems can either store a copy of the content or retrieve the content from the KCMS on demand via an application programming interface (API). Knowledge implementation
systems execute actionable knowledge artifacts, such as alerts and reminders or order sets. The knowledge implementation systems generally store the executable version of knowledge artifacts, built using proprietary languages and information models, but could request an executable knowledge artifact from the KCMS if it is built in a standard executable format. A high-level overview of the KCMS architecture is illustrated in Figure 5.

Managing knowledge in a dedicated system requires an ontology of knowledge assets which clearly defines the various asset types, the metadata used to describe the assets, and their relationships. While a working list of asset types has been developed to meet local needs, a broadly accepted standard will be necessary to effectively enable the collaboration and exchange across organizations. The knowledge asset ontology describes what an asset is (content, a rule, a predictive model, an order set, etc.), but annotation with standard medical vocabularies (e.g. SNOMED-CT, LOINC®, ICD10-CM, and RxNorm) is required to describe what the asset is about. Four internally-developed vocabularies are used to support specialized uses, such as consumer health and research. Early experience with annotation of assets showed that while it was useful in understanding broad relationships between assets, it was not sufficient to describe the relationships between specific assets. Therefore, a mechanism was developed to allow declaration of explicit relationships between knowledge assets (Figure 4). For example, the treatment section of a care process model might recommend medications and therapies for a particular condition. The order set derived from that section would have an explicit relationship to the section, so that when the section is updated, notification goes to the order set team for review and update as needed. These relationships quickly grew in complexity; for example, a rule derived from a care process model might depend on a scoring model, which might depend on a cohort definition, which might depend on a value set. At this time, the types of item relationships created and maintained in the KCMS are limited to “dependencies” and serve to notify the owner of an asset about changes to related assets. Human intervention is required to analyze and react to the changes. Because the ability to define explicit relationships is a fairly recent enhancement, and requires business processes that cross multiple teams, the number of such relationships is relatively small but growing; as of this writing, there were 3131 relationships between items, 792 of which were subscriptions to changes in content. In the future, a more complete asset ontology that incorporates the medical semantics of asset to asset relationships is obviously desirable.
Integrating the KCMS with the EHR also poses challenges. While many of the assets themselves are built in the EHR, the knowledge about the asset, including descriptive and business metadata, and the relationship to other assets within and outside the EHR, are managed in the KCMS. The current mechanism for integration is cross-referencing. The KCMS record of a knowledge asset contains the unique identifier from the EHR configuration record, and the EHR configuration record contains the unique identifier of the KCMS catalog. Our intent is to leverage this cross-referencing to allow a user of the KCMS to view the EHR implementation details from within the catalog, and to allow a rule builder in the EHR to view the system-agnostic rule logic, the metadata, and relationships to other knowledge assets from within the EHR.

The motivation for this approach to knowledge management reflects both structural and operational goals. Structurally, we want to reflect an ideal conceptual model which requires the independence of clinical knowledge representing the standard of care at Mayo Clinic from the clinical workflow embodied in the EHR’s programming. Operationally, the design has to satisfy our need to “know what we know” – that is, to allow a Knowledge Content Board to efficiently access the entire body of knowledge related to a given condition, clinical situation, or treatment in order to validate the clinical relevance and appropriateness of the various knowledge artifacts and relationships, and to systematically update them as our recommended standard of care evolves.

Example: Updating Recommendations for Anticoagulation and Atrial Fibrillation

One of Mayo's CPMs addresses recommendations for anticoagulation in patients with atrial fibrillation. Several rules in the EHR and other systems calculate a patient’s risk of stroke and make recommendations regarding anticoagulation therapy. In 2014, the clinical guidelines were updated in response to rapidly developing continuous change in assessment and treatment, based upon the results of multiple randomized trials being conducted to determine the optimal therapy to decrease the risk of stroke in patients with atrial fibrillation. Examples of the changes include a) the introduction of the new oral anticoagulant medications with a superior effect and better patient compliance than warfarin, b) the creation of better models for risk prediction (i.e., the CHADS2-VASC vs CHADS risk score calculators), c) the deletion of treatments that had been passed down as effective but further analysis showing ineffectiveness (i.e. aspirin), and d) the need for a shared decision making process with the patient in which they make a decision based upon their own needs and preferences. After the new evidence was reviewed by clinical experts and a consensus was reached, Mayo staff was able to update the related rules in the KCMS quickly and efficiently, thereby ensuring that clinicians across the Mayo enterprise had access to the new information. Outdated information persisted in isolated reference documents on the intranet and on small departmental systems, however, and these sites were updated as they were identified. This example provided a tangible demonstration of the value of a formalized, semi-automated knowledge management approach.

Evaluation

In October 2014, Mayo hosted three renowned clinician-informaticians to conduct an external review of the Mayo Clinic Knowledge Management Program. The key questions posed to the external evaluation team included:

- What are the best practices for managing clinical guidelines and pathways?
  - Knowledge structure (including degree of granularity and relationships between topics)
  - Knowledge representation (from human-readable to executable)
  - Tools used to store and manage knowledge
  - Methods for indexing and retrieval (e.g., annotations, tags)
  - Knowledge delivery
- How should other knowledge assets be managed?
  - Types of assets (e.g., order sets, clinical decision support rules, educational materials)
  - Relationships between knowledge assets (if any)

The external evaluation team noted the breadth and depth of the information maintained, curated, and distributed on a daily basis at Mayo Clinic. The group also approved of Mayo’s use of clinical ontologies to act as building blocks of the information architecture, citing the importance of incorporating standards into the process and applying appropriate and usable semantic tags.

As with many large academic medical centers, Mayo’s knowledge asset development and curation process is necessarily dispersed throughout the practice. The team recommended the assets be managed by subject matter, regardless of the delivery system intended. This approach will require a greater level of coordination from the knowledge management program, but is necessary to ensure the knowledge delivered to multiple sources is created and vetted by the appropriate subject matter experts. The evaluation team also identified opportunities for
improvement in coordinating of multiple delivery systems, particularly based on target audiences of providers, patients, consumers and affiliated practices.

The external evaluation team commented favorably on Mayo’s information model and architecture for the technical management of clinical knowledge. The team encouraged Mayo to advance the technical capabilities to facilitate the ability of subject matter experts to use the central repository of knowledge assets, which will reduce errors in transcription from one tool to another and encourage a more complete and focused entry of the assets. In addition, it will reduce the likelihood of the assets stored in local repositories, which enforces the centralized source of truth. Finally, the team noted that many of the knowledge assets will ultimately be delivered through the electronic health record (EHR), and those assets (e.g., order sets, decision support rules, and representations of clinical pathways) should have a clear pathway between the central repository of knowledge and the EHR.

As shown in Table 2, the KCMS was also evaluated against the recommendations outlined by Sittig et. al. While those recommendations focus specifically on systems intended to manage clinical decision support knowledge, the principles apply more broadly to clinical knowledge management.

<table>
<thead>
<tr>
<th>Tool 1: external repository of clinical content with web-based viewer</th>
<th>The KCMS is an external repository many types of knowledge assets. Assets can be discovered by end users through a variety of interfaces and applications. Developers and knowledge management staff can access the computable knowledge assets via the Knowledge Asset Catalog web page.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tool 2: online, collaborative, interactive, Intranet-based tool to facilitate content development</td>
<td>The KCMS supports multiple workflows for collaboration, review, and approval and captures metadata about the knowledge asset life cycle. However, the workflows are not always intuitive and users express dissatisfaction with current processes. Enhancements are currently being made to the system to improve the experience. Multiple authoring tools for various asset types are being evaluated.</td>
</tr>
<tr>
<td>Tool 3: enterprise-wide tool to maintain controlled clinical terminology concepts</td>
<td>Mayo Clinic’s “TERMS” environment manages standard and home-grown terminologies and value sets. Currently, the rules must reference a terminology web page that lists the approved value sets, but we are working toward each value set having a URI to enable direct referencing from within the rule representation.</td>
</tr>
<tr>
<td>Tool 4: tool for CDS users to provide feedback regarding specific CDS interventions</td>
<td>There is no single tool for users to provide feedback on CDS; the alerts and reminders include links to email the rules team, AME include a feedback window which stores the feedback in a local database. The feedback is used to improve the program, but it is not managed as systematically as it could be.</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of KCMS using principles of clinical knowledge management

Future Work

Although much progress has been made in leveraging a core body of clinical knowledge as an anchor for managing multiple derivative knowledge assets, Mayo Clinic continues to actively address the significant opportunities and challenges. This activity will be done in collaboration with other academic medical centers and healthcare experts to expand capabilities for all developers and consumers of clinical knowledge.

The current model of extracting executable logic from CPMs is an improvement over starting from a blank slate with a subject matter expert, but the process remains labor-intensive and requires significant effort to manually update the derived assets when the underlying knowledge changes. More exploration is needed in approaches to modeling the core knowledge in a way that streamlines the generation of executable derivatives while maintaining the value of the narrative representations.

Managing cohort definitions, scoring models, and other inputs to clinical rules as independent assets enables reuse and enhances scalability, but better mechanisms are needed to name, classify, and relate these knowledge assets. As a starting point, Mayo will lead a collaborative effort to develop a common ontology of knowledge assets. Agreement on clear definitions of knowledge assets and their relationships to each other is required to facilitate sharing of any complex representation of knowledge.

While there are many proposed representation formats and standards, they may not be uniformly adopted, may not be sufficiently expressive (e.g., Health eDecisions, HeD), or may lack sufficient specificity (e.g., HL7 FHIR) so...
that true exchange of computable, actionable knowledge is still a challenge. Ongoing collaboration between organizations, with an emphasis toward applying standards and models in practice, is critical to advancing knowledge management.

Further work is also needed to bring managed knowledge assets together with patient data within clinical workflow in a fashion that efficiently assists and supports the cognitive effort of clinicians using the knowledge assets. This requires precise alignment of the concepts in the knowledge assets with the data schema of the EHR, and a careful elaboration of how clinicians seek and interact with knowledge in the natural flow of diagnosis and treatment decisions.

Another facet of knowledge management which merits further exploration is knowledge discovery through measurement. The knowledge lifecycle depends on measuring the results of knowledge delivery and using that data to feed back into knowledge creation (Figure 6). Capturing data as a byproduct of knowledge delivery interventions in the EHR creates a rich pool of data that can be used to measure and analyze practice variation. This data can serve as an important mechanism for continuously improving the CPMs\textsuperscript{13}. Looking for ways to identify the data that will be used to measure the effectiveness of a knowledge delivery intervention as a part of the knowledge engineering process will become increasingly important in the evolution of learning health systems.

Figure 6. The knowledge lifecycle

Conclusion

Physicians make decisions that must be grounded in expert clinical knowledge almost continuously during the process of caring for patients. Mayo Clinic developed a highly collaborative model of care based on a unified patient record starting early in the 20\textsuperscript{th} century, and this developed over time into a system of both formal and informal knowledge sharing directly from physician to physician as the primary means of assuring access to clinical knowledge both broad and deep enough to facilitate optimal care decisions. In the 21\textsuperscript{st} century era of the EHR, Mayo has created a digital knowledge sharing system as a complement to the EHR, designed to allow clinicians direct electronic access to clinical knowledge that represents the Mayo Clinic standard of care for 1300 clinical conditions. This system has required elaboration of a formal system for knowledge development and stewardship as part of the medical responsibilities of Mayo physicians, and resulted in the construction of a Knowledge Content
Management System used to organize and deliver knowledge on demand in clinical and other settings. Full utilization of ontological knowledge management and integration of knowledge with patient data maintained in the EHR remain challenges and areas of ongoing development. While there are several national leaders in this area we believe that our approach to clinical knowledge management, which is centered on structured, vetted content that is used as the basis for a variety of derived assets, is unique and can serve as a model for other enterprise knowledge management efforts.

Acknowledgements

The authors thank Matt Burton, Scott Eggers, Michelle Felten, Ravi Nadimpally, Josh Pankratz, Kris Pavek, Kristi Sorensen, Davide Sottara, and Jocelyn Webb for assistance in the preparation of this manuscript. We are also extremely grateful for the insight provided by James E. Tcheng, William R. Hersh, and Kensaku Kawamoto, who conducted the external review of the Mayo Clinic Knowledge Management Program.

* J. Shellum and R. Freimuth contributed equally to this work.

References

8. T., S: Can you explain that again? DITA for beginners.
Automatic data source identification for clinical trial eligibility criteria resolution

Chaitanya Shivade, MS\textsuperscript{1}, Courtney Hebert, MS, MD\textsuperscript{2}, Kelly Regan, BA\textsuperscript{2}, Eric Fosler-Lussier, PhD\textsuperscript{1}, Albert M. Lai, PhD\textsuperscript{2,3}

\textsuperscript{1}Department of Computer Science and Engineering
\textsuperscript{2}Department of Biomedical Informatics, The Ohio State University, Columbus, OH.
\textsuperscript{3}National Institute of Health, Rehabilitation Medicine Department, Mark O. Hatfield Clinical Research Center, Bethesda, MD.

Abstract

Clinical trial coordinators refer to both structured and unstructured sources of data when evaluating a subject for eligibility. While some eligibility criteria can be resolved using structured data, some require manual review of clinical notes. An important step in automating the trial screening process is to be able to identify the right data source for resolving each criterion. In this work, we discuss the creation of an eligibility criteria dataset for clinical trials for patients with two disparate diseases, annotated with the preferred data source for each criterion (i.e., structured or unstructured) by annotators with medical training. The dataset includes 50 heart-failure trials with a total of 766 eligibility criteria and 50 trials for chronic lymphocytic leukemia (CLL) with 677 criteria. Further, we developed machine learning models to predict the preferred data source: kernel methods outperform simpler learning models when used with a combination of lexical, syntactic, semantic, and surface features. Evaluation of these models indicates that the performance is consistent across data from both diagnoses, indicating generalizability of our method. Our findings are an important step towards ongoing efforts for automation of clinical trial screening.

Introduction

With the amount of data collected during patient care growing, researchers have proposed a number of scenarios in which knowledge extracted from clinical data can be leveraged effectively\textsuperscript{1}. Information found in the Electronic Health Record (EHR) can have an important secondary use in identifying patients eligible for clinical trials\textsuperscript{2}. Identifying the patient cohort involves understanding the exact information sought by the eligibility criteria for trials and searching for it in the subject’s EHR. However, this manual process of eligibility determination is challenging and time-consuming\textsuperscript{3}. Therefore, a number of initiatives have been undertaken to introduce automation into this process\textsuperscript{4}.

The EHR captures patient information using both structured data, stored in the form of tables or flowsheets, and unstructured data, stored in the form of clinical notes or free-text reports. While structured data are comprised of laboratory values, demographics, diagnosis codes, etc., unstructured clinical notes are written by physicians, nurses and other health professionals to document a patient’s condition, prognosis, response to various clinical interventions, and future plans for treatment. Eligibility criteria based on data found in structured data fields can be processed easily using structured queries over databases (e.g., “At least 18 years of age at the time of enrollment.”). Although structured data enables quick and easy retrieval of data, unstructured text is a preferred means of documentation for physicians\textsuperscript{5}. Thus, the information need of eligibility criteria that involve clinical nuances (e.g., “Patients with a known severe symptomatic primary pulmonary disease.”) can only be answered using unstructured clinical notes or discussion with the treating physician.

Gaps between data found in the EHR and the desired data for trial recruitment have been studied\textsuperscript{6}. Köpcke et al.\textsuperscript{7} found that there was a significant gap (65\%) between the structured data documented for patient care and the data required for eligibility assessment. Li et al.\textsuperscript{8} studied the strengths and weaknesses of using structured or textual clinical data for clinical trials eligibility screening. They concluded that insights obtained from natural language processing of patient reports supplement important information for eligibility screening and should be used in combination with structured data. Therefore, researchers have devised methods\textsuperscript{9,10} using NLP to identify relevant text in clinical notes and resolve eligibility criteria to screen patients into clinical trials. However, using NLP techniques on clinical data is computationally expensive. Moreover, certain criteria (e.g., “Patients between the ages of 18 and 90.”) can be answered accurately and quickly using structured data alone. An ideal system for identifying
patients would maximize the use of structured data, and only use inference over clinical narratives when necessary. Thus, it would be desirable to know whether eligibility can be concluded for a criterion using structured data, or if it requires an NLP pipeline for its resolution. Such knowledge would not only result in identifying patient cohorts with better accuracy but also optimize the computational efficiency of the task.

Related Work

Eligibility criteria for clinical trials are specified in natural language and hence they are not amenable to computational processing. In their comprehensive review of formal representations of eligibility criteria, Weng et al.\(^\text{11}\) discuss the variety of intermediate representations used by researchers to overcome this challenge. The authors discuss classification of eligibility criteria along three dimensions: content, use in eligibility determination, and complexities of semantic patterns. The first dimension deals with the information needed to answer eligibility queries. Multiple studies\(^\text{12,13,14}\) have been published characterizing this dimension of eligibility criteria. The TrialBank project\(^\text{14}\) classifies criteria into three large categories: age-gender, ethnicity-language, and clinical. Tu et al.\(^\text{15}\) classified criteria from a clinical perspective: stable, variable, controllable, subjective, and special. Metz et al.\(^\text{16}\) classified criteria for cancer trials as demographics, contact information, personal medical history, cancer diagnosis, and treatment to date. The majority of these classifications are very specific to the downstream task of these studies. Moreover, since most of these are annotated manually, they are also time-consuming and expensive.

Very few studies have explored automatic classification of eligibility criteria. Luo et al.\(^\text{17}\) classified eligibility criteria into 27 semantic categories. They show that machine learning classifiers using semantic types from the Unified Medical Language System (UMLS) as features perform better than those using traditional bag of words features; with J48 decision trees performed better than Bayesian models. Bhattacharya and Cantor\(^\text{18}\) present a study where they automatically map trial protocol documents from the pharmaceutical industry to openly available trial specifications. They compare the eligibility criteria from both sources and propose the creation of template criteria for standardization. Levy-Fix et al.\(^\text{19}\) propose automatic mapping of eligibility criteria to the widely adopted Common Data Model using a clustering method that uses UMLS semantic type based features. Hong et al.\(^\text{20}\) define “diagnostic criteria” as a combination of signs, symptoms, and test results used by clinicians to determine the correct diagnosis, and propose mapping of such criteria to the Quality Data Model.

In this paper, we target a very specific problem of automatically classifying whether an eligibility criterion should be resolved using structured data or by using NLP techniques. This is an application driven problem that is common to a number of approaches undertaken by the biomedical informatics community for the secondary use of EHR data. We present a detailed evaluation of our approach using data from two disparate diagnoses. The paper is organized as follows: in order to evaluate an approach that can automatically classify eligibility criteria, we first describe the creation of a gold standard dataset. This is followed by a description of the features used for the machine-learning based classification task and our results. We then examine errors made by our system, followed by a discussion and conclusion of our work.

Dataset Creation

The National Library of Medicine and the National Institutes of Health maintain the website www.clinicaltrials.gov, a registry and a database of publicly and privately supported clinical studies across the globe. Almost all of the studies described above, which analyze eligibility criteria of clinical trials, use this registry as their data source. We searched the registry with the query “heart failure” and randomly sampled 50 trials from the result. This was repeated for the query “chronic lymphocytic leukemia.” We then extracted all eligibility criteria from these 100 trials and split them into individual criterion sentences using a sentence chunking module from Lingpipe\(^\text{21}\) that was trained on MEDLINE articles, along with custom rules for refinement. The goal was to annotate each criterion with the preferred data source for resolution, either “Structured” or “Unstructured.” We selected trials from two disparate diseases to test the generalization capabilities of our method. The goal was to verify if preferences for resolving these criteria would remain consistent across trials of different diagnoses.

Our annotation team consisted of two senior undergraduate nursing students, a medical student (KR), and a physician (CH). The first author (CS) developed detailed guidelines for the annotators in consultation with the physician on the team. While creating the annotation guidelines, we assumed that the criteria were being resolved by searching for patient information in the EHR software from Epic Systems (Verona, WI) at our institution, The Ohio State University’s Wexner Medical Center. In order to ensure that the annotation guidelines were complete, we
conducted three training rounds with ten trials each. This was followed by a discussion period, and disagreements were resolved by the physician. The annotation guidelines were updated after every training round to reduce differences among annotators. We ensured that trials used in the training rounds were not included in the final set of trials used for the study. The training rounds consisted of 30 trials for annotation; criteria texts and annotations from these trials were then used for feature development.

During the discussions, our annotators argued that although the EHR might have a structured field for a criterion, the field may or may not be consistently used by providers for documentation. We addressed this by assuming that for a particular criterion, if the EHR had an available structured field for documentation, it would be classified as structured. In cases of criteria where evidence from both structured and unstructured sources was required, we treated the annotation as unstructured, since criterion resolution could not proceed unless a clinical note was referred to. The end-goal of the classifier being developed is to separate eligibility criteria that can be resolved exclusively by structured data from others, in order to speed up screening. Thus, if a criterion can be partially resolved using structured data but mandates reading of notes for its full resolution, it cannot be processed exclusively using structured data and hence does not contribute to speedup. We therefore do not consider a separate category for “Structured AND Unstructured.”

Another observation made by our annotation team was the presence of eligibility criteria in trials that could not be resolved by information in the EHR at all. These included criteria that would need to be resolved in person or after the start of the study, such as informed consent, willingness to comply with certain restrictions, and criteria associated with post-recruitment details such as randomization constraints, etc. Therefore, in addition to the two categories of structured and unstructured data, we included a third category “Neither” for our annotations. We also observed that the automatic sentence segmentation was not accurate; these were hand-corrected by our annotators. The final kappa among annotators on the set of 50 heart failure trials was $\kappa = 0.83$, with $\kappa = 0.81$ for the CLL trials; these indicate excellent inter-annotator agreement. The dataset consists of 766 criteria across 50 heart failure trials and 677 criteria across 50 CLL trials. Table 1 summarizes the distribution of annotations across the three categories, which is fairly consistent across the two diagnoses. Figure 1 shows sample criteria across each of the three categories.

### Table 1. Overview of annotations in the dataset

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Number of criteria</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HF</td>
<td>CLL</td>
</tr>
<tr>
<td>Unstructured</td>
<td>491</td>
<td>368</td>
</tr>
<tr>
<td>Structured</td>
<td>194</td>
<td>193</td>
</tr>
<tr>
<td>Neither</td>
<td>81</td>
<td>116</td>
</tr>
</tbody>
</table>

**Figure 1.** Sample eligibility criteria across the three classification categories.

**Unstructured**
- Major heart event or heart procedure within the 6 weeks prior to study entry.
- Stable, chronic, mild to severe heart failure as defined as subjects with symptoms of heart failure who do not require IV diuretics, inotropes, or vasodilators or those that require support with a left ventricular assist device.
- Progressive lymphocytosis with an increase of > 50% over a 2 month period, or an anticipated doubling time of < 6 months.

**Structured**
- porphyria
- Individuals who are morbidly obese
- Has significant, moderate-severe renal dysfunction or disease, confirmed by serum creatinine of greater than 2mg/dl (180 micromol/L)

**Neither**
- Patients must be able to provide informed consent.
- Concomitant administration of oral ACE inhibitor medication from 2 hours before the start of the study drug until 30 minutes after the start of the study drug.
- Patient is physically capable and willing to perform repeated physically demanding tests associated with the study.
Feature Representation

We extracted four sets of features for creating a model to classify the eligibility criteria statements. As discussed earlier, these were designed by examining the criteria from the training rounds. We used MetaMap\textsuperscript{22} for all UMLS related features and the Stanford Parser\textsuperscript{23} for other shallow linguistic features. We explain each feature and the rationale for its design below:

Surface Features
M1. Number of words (excluding stop words): Short criteria can be often inferred from structured data.
M2. Number of numerical quantities: Presence of numerical quantities reduces ambiguity and can thus be inferred from structured data.
M3. Number of arithmetic symbols: Presence of arithmetic symbols such as ‘<’, ‘>’ or ‘=’ typically indicates a criterion that can be easily converted into a structured query.
M4. Number of delimiters: Protocol writers often combine multiple sub-criteria using a semi-colon or comma. Multiple information sources may be required to resolve such a criterion.
M5. Number of measurement units: These are often lab values and can be reliably inferred from structured data.
M6. Number of UMLS concepts: This has the same intuition as M1. Criteria with multiple concepts are likely to be complex and thus would mandate reading of notes.
M7. Number of illustrative terms: A generic eligibility criterion is often explained using examples. Illustrative terms used for such a construction are: \textit{e.g.}, \textit{example}, \textit{such as}, \textit{defined by}.

Syntactic Features
SY1. Number of conjunctions: A single criterion often has multiple sub-criteria in it separated by conjunctions (such as \textit{and, or}). A criterion with multiple sub-criteria is likely to be resolved by reading notes.
SY2. Number of adjectives: Adjectives are used to describe qualitative judgments, possibly making the criterion ambiguous from a computational perspective. It is hard to conclude about such criteria using structured data.

Semantic Features
SM. Semantic types of concepts: Semantic types from UMLS provide a level of abstraction that is intuitive for criteria classification. For example, it is generally accepted that criteria involving laboratory values and medications are resolved using structured data. These correspond to well-defined UMLS semantic types “Laboratory or Test Result” and “Pharmacological Substance” respectively.

Lexical Features
L1. Bag of Words: As discussed earlier, all criteria are from trials of specific diagnoses. Thus, similar words will be commonly used for defining criteria from each of the three categories. These features will capture this pattern.
L2. Bag of UMLS Concepts: This has the same rationale as L1 with one difference: UMLS concepts are often multi-word expressions, and a concept level mapping helps to mitigate this phenomenon.

Results

We evaluated the contributions of different features for the criteria classification task using machine learning classifiers. The evaluation of these tasks was carried out using accuracy as a metric. The distribution of the three categories in our data shows that the simplest classifier, which always predicts the category “Unstructured” for every instance, can achieve an accuracy of 64% on heart failure and 54.4% on CLL. This “Majority” classifier is the simplest baseline.

We chose two different machine learning algorithms for our experiments, namely Naïve Bayes and Support Vector Machines (SVMs). While Naïve Bayes is a simple classifier, it can deliver quick results. SVMs are state of the art machine learning algorithms for classification tasks and work well even with datasets that are not linearly separable. Although it can be time consuming to tune SVM parameters on large datasets, the relatively small dataset in our work meant tuning was not an issue. Following Luo et al.,\textsuperscript{17} we also tried J48 decision trees, but did not observe any significant performance gain in comparison with the Naïve Bayes model.

We set up the classification task in two settings. First, we carried out a leave-one-out validation on a per-trial basis. Of the 50 trials in our dataset, we trained a classifier on eligibility criteria across 49 trials and tested it on the
remaining trial. This was repeated 50 times, such that each trial appeared in the test set once, to calculate the reported accuracy. We refer to this as the trial-dependent setup. Second, we carried out criteria classification independent of their association with a trial. Each criterion statement was considered as a separate instance and evaluation was performed using ten-fold cross validation. While such a setup may induce some knowledge into the model since criteria from the same trial may appear in the training set, it is a more practical setup. When such a system will be deployed for practical use, it would be trained on criteria from different trials. The results of these two evaluations are summarized in Table 2. We found that SVMs outperformed Naïve Bayes significantly; interestingly, the trial-independent setup gave better results with both models than the trial-dependent one.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Naïve Bayes</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial dependent</td>
<td>HF 70.4</td>
<td>CLL 71.7</td>
</tr>
<tr>
<td></td>
<td>HF 75.2</td>
<td>CLL 77.7</td>
</tr>
<tr>
<td>Trial independent</td>
<td>HF 71.3</td>
<td>CLL 74.2</td>
</tr>
<tr>
<td></td>
<td>HF 79.8</td>
<td>CLL 82.1</td>
</tr>
</tbody>
</table>

**Feature Analysis**

The Naïve Bayes classifier assumes that each of the features it uses is conditionally independent of the others, given some class label. SVMs do not make this assumption. They transform a given set of data points into a high dimensional space and construct a hyperplane for classification. We hypothesized that the Naïve Bayes model does not perform well because it does not consider the interaction between different features. In order to verify this, we evaluated the contribution of our various features towards the performance of the two models, using the trial independent setup. We compare feature contributions for both the Naïve Bayes and SVM classifiers. The following sections discuss experiments for these evaluations.

**Naïve Bayes**

We created simple classification models using the Naïve Bayes classifier in combination with the L1, L2, and SM feature sets individually. A fourth model was also created combining all the morphological and syntactic features. All models were created using the Weka toolkit. The results of these individual models are summarized in Table 3.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Constituents</th>
<th>Number of features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>-</td>
<td>HF 64.1</td>
<td>HC 54.4</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>L1</td>
<td>HF 1874</td>
<td>CLL 1616</td>
</tr>
<tr>
<td>Bag of Concepts</td>
<td>L2</td>
<td>HF 2246</td>
<td>CLL 1815</td>
</tr>
<tr>
<td>Bag of Semantic Types</td>
<td>SM</td>
<td>HF 101</td>
<td>CLL 102</td>
</tr>
<tr>
<td>Morphology and Syntax</td>
<td>M1-M5, SY1-SY2</td>
<td>HF 9</td>
<td>CLL 9</td>
</tr>
<tr>
<td>All</td>
<td>M, L, SY, SM</td>
<td>HF 4230</td>
<td>CLL 3542</td>
</tr>
</tbody>
</table>

**Support Vector Machines**

SVMs belong to a class of algorithms known as kernel methods, which have been shown to be very effective for classification. This is achieved through a kernel function, which is a similarity function $K$ that maps two inputs $x$ and $y$ from a given domain into a similarity score that is a real number. Formally, it is a function $K(x, y) = \langle \phi(x), \phi(y) \rangle \rightarrow \mathcal{R}$, where $\phi(x)$ is some feature function over instance $x$. For a function to be a valid kernel, it should be symmetric and positive semi-definite.

To analyze the contribution of different features, we constructed individual linear kernels (regularization parameter $C=0.25$) using each of the four feature sets described earlier: words, concepts, semantic types, and morphology and...
syntax features. This involves creating a kernel matrix (or gram matrix) comprised of similarity values between all pairs of instances in the dataset, obtained using the kernel function \( K \). All models were implemented using the LibSVM library. Joachims et al. also showed that given two kernels \( K_1 \) and \( K_2 \), the composite kernel \( K(x, y) = K_1(x, y) + K_2(x, y) \) is also a valid kernel. This can be achieved by a simple addition of two kernel matrices. Thus, we combined different kernels to test the contribution of individual feature sets towards classification. The results are summarized in Table 4.

Comparing results from Table 3 and Table 4, we observe that SVMs perform better at the classification task by modeling interactions between different features. In the following section we discuss the performance of this classifier across each of the class labels: “Structured”, “Unstructured” and “Neither”. We also discuss common errors made by the classifier in this task.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Features</th>
<th>HF</th>
<th>CLL</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>-</td>
<td>64.0</td>
<td>54.4</td>
<td>59.1</td>
</tr>
<tr>
<td>Words</td>
<td>L1</td>
<td>76.2</td>
<td>76.8</td>
<td>74.5</td>
</tr>
<tr>
<td>Concepts</td>
<td>L2</td>
<td>75.0</td>
<td>76.8</td>
<td>75.7</td>
</tr>
<tr>
<td>Semantic Types</td>
<td>SM</td>
<td>74.1</td>
<td>70.0</td>
<td>69.6</td>
</tr>
<tr>
<td>Morphology and Syntax</td>
<td>M1-M5, SY1-SY2</td>
<td>66.1</td>
<td>55.0</td>
<td>61.2</td>
</tr>
<tr>
<td>All</td>
<td>L, SM, M, SY</td>
<td><strong>79.8</strong></td>
<td><strong>82.1</strong></td>
<td><strong>80.6</strong></td>
</tr>
</tbody>
</table>

**Error Analysis**

As discussed in the section on data creation, the corpus is skewed, with the majority of the eligibility criteria labeled as resolved from unstructured data. Table 5 presents the confusion matrix of the SVM classifier (using the “All” kernel combination from Table 4) on the combined data. These errors are representative of the individual datasets as well. Most of the errors stem from the class “Structured,” while the “Unstructured” class has the least.

<table>
<thead>
<tr>
<th>Label</th>
<th>Prediction</th>
<th>Number of instances</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured</td>
<td>768</td>
<td>859</td>
<td>89.4</td>
</tr>
<tr>
<td>Structured</td>
<td>129</td>
<td>387</td>
<td>64.8</td>
</tr>
<tr>
<td>Neither</td>
<td>46</td>
<td>197</td>
<td>73.1</td>
</tr>
</tbody>
</table>

We examined the criteria for these errors. Many criteria are simply too brief and the classifier has very little context to predict the right label. For example, the criteria “Any etiology” or “Diabetes Mellitus” are labeled as “Structured” but misclassified as “Unstructured.” Similarly, “Previous exposure to KW-3902” or “End stage cancer” are labeled as “Unstructured” but predicted as “Structured.”

Interestingly, some errors also stemmed from information that would be classified as “semi-structured” data by humans. These were discussed by the annotators during our training rounds for consensus. For example, one of the guidelines was to label criteria associated with numerical fields found in ECG and EKG data as “Unstructured,” since they are not present as discrete values in our EHR. The classifier made errors for statements such as “EF < 45 %” or “QRS > 120 msec.” Similar observations can also be made for criteria such as “Folstein MMSE score less than 20” or “Walk < 450 meters during 6 minute hall walk test.”

Although certain criteria statements had a single class label assigned to them, there was a possibility of multiple labels being associated with them. For example, consider the criterion statement: “Adult patients age >= 18 admitted with CHF exacerbation with NYHA Class III-IV symptoms at screening.” The annotators marked this criterion as
“Unstructured,” whereas our model classified it as “Structured.” One can observe that this statement has two sub-criteria: one related to the patient being adult, and the other one about her CHF. While the age related part can be resolved using structured data, the CHF part requires reading clinical notes. A similar claim can be made for “WBC > 1.5; ANC > 500; Plt > 50,000 unless documented as due to disease” which was marked “Unstructured” by our annotators (because of the judgment required to determine if these laboratory abnormalities were related to the disease) and predicted “Structured” by our model.

Some kinds of criteria were quite rare and hence the model was not able to predict the correct label. For example, the criterion “Currently a prisoner” was labeled as “Structured” but predicted as “Unstructured.” This is the sole criterion about prisoners in our dataset. Similarly, “Inability to lie flat for MR study” is labeled as “Neither” and misclassified as “Unstructured.” There were some obvious generalization errors as well. For example, “Current treatment on any Class I or III antiarrhythmic, except amiodarone,” which was labeled as “Structured” but incorrectly predicted as “Unstructured.” In the trials for heart failure, mentions of “Class” are associated with New York Heart Association (NYHA) Class I, II, III or IV and are labeled as “Unstructured” multiple times in the dataset. Similarly, for the criterion “Patients who are unable to refrain from taking acetaminophen” the phrase “unable to refrain” makes it “Unstructured,” however the remaining is a medication that can be looked up in the structured medication list, so our model classified it as “Structured.” A larger labeled dataset or lexical features with longer n-grams could potentially avoid such errors.

Some errors could not be explained. For example, “Patients having undergone revascularization procedures within 6 months” or “Current treatment of calcium channel blockers except for long acting dihydropyridines” were labeled as “Structured” and misclassified as “Unstructured.” Similarly, “End-stage renal disease on dialysis or imminent” was labeled as “Unstructured” (would not be able to determine “imminent” without reading a note) and misclassified as “Structured.”

**Discussion**

Figure 2 outlines the role of our work in an automated clinical trial screening workflow. A number of clinical research studies mandate selecting patients as subjects based on stringent eligibility criteria. Clinical trials are a particular type of such a research study. These eligibility criteria are specified in natural language text (block 1), and trial coordinators examine these criteria to locate information in the EHR (block 2) that can be used to resolve the criteria. This information is present in the form of structured content such as tables and flowsheets (block 2.1), or in the form of clinical notes and reports (block 2.2). Eligibility decisions (block 3) are made based on evidence from these sources.

**Figure 2.** Role of this work in an automated clinical trial screening workflow
There has been a lot of research focused around introducing automation into the process of cohort identification, with various efforts in the research community addressing different stages in the screening process. Recent studies in NLP focus on the development of techniques that can be used to automate cohort identification. These can assist the human effort involved in reading the unstructured content from the EHR for eligibility determination (block 4). A number of studies in the past have proved that structured information such as billing information can be used effectively for cohort identification. These are stored as structured content in the EHR in the form of databases. Simple SQL queries can be used to extract this information (block 5). There has been a large body of work to translate eligibility criteria into such queries. However, to the best of our knowledge, there has been no work in the past to triage an eligibility criterion along one of these two routes. This work is an effort towards introducing automation in this step of the workflow. This work has utility in various applications such as clinical question-answering and decision support where the same problem of choosing a preferred data source for evidence gathering is encountered.

Limitations

This is an initial investigation for a potentially larger problem and is not without limitations. As discussed in the section on dataset creation, we made some assumptions for the task of annotation. Our most crucial assumption was the availability of data in a structured field; in practice, presence of a structured field in the EHR does not imply availability and reliability of the values in that field. Retrospectively, we realized that although the annotation team had high agreement, some decisions were subjective based on the annotators’ knowledge of our EHR, and could differ at other institutions. In some cases, the criterion may be best resolved by looking at both structured and unstructured data which cannot be sped up despite the classification step using our model. Finally, annotations obtained for this study were based only on observing and reading the criteria texts. A retrospective study examining agreement for our annotations with actual screening undertaken by clinical trial coordinators would confirm our findings.

Conclusion

We introduced the problem of identifying the right data source type for resolving an eligibility criterion in clinical trial screening. A gold standard dataset for automating the task using 100 trials (50 for heart failure, and 50 for CLL) was created. We focused on the language used in the criterion text and designed a variety of intuitive features for determining which data source to use to best resolve a particular criterion. These include morphological, lexical, syntactic and semantic features. An empirical comparison of two popular machine learning algorithms shows that a combination of these features results in optimal performance using SVM as a classifier. We note that most of our remaining errors stem from structured criteria being predicted as unstructured. Our system’s current performance shows promise, but has room for improvement. Our work is a step towards efficient automation of clinical trial recruitment by triaging each eligibility criterion to the right data source for its resolution.

Acknowledgements

We would like to thank Jennifer Fox and Christy Starr for their help with the annotation effort. We would also like to thank Denis Griffis for his help in preparing in this paper. Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011116, the Intramural Research Program of the National Institutes of Health, Clinical Research Center, and through an Inter-Agency Agreement with the US Social Security Administration. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


Better managing technology-mediated interruptions in the ICU: Examining the role of patient information for improving text message notifications

Preethi Srinivas, PhD1, Madhu C. Reddy, PhD, FACMI2, Anthony Faiola, PhD3
1Indiana University School of Informatics and Computing, Indianapolis, IN; 2Northwestern University, Evanston, IL; 3University of Illinois, Chicago, IL

Abstract
Previous research has identified the need for managing wanted and unwanted interruptions from technology-mediated notifications (TMN) in the intensive care units (ICUs). Current solutions are focused on mobile, asynchronous and context-aware mechanisms that consider a minimal number of factors (location and activity of the receiver). These factors are insufficient for a receiver to effectively decide on whether or not to interrupt their ongoing activities to immediately respond to a TMN. We propose a mobile device solution, known as “patient-enhanced notifications” that presents a preview of TMN with additional patient information. A study comprising of user evaluations and interview sessions helped ascertain that patient vital signs coupled with the actual text message assisted receiving ICU providers in deciding on when to respond to the TMN. We conclude that patient-enhanced notifications has the potential to help ICU clinicians better manage interruptions generated from mobile devices.

Introduction
Hospital intensive care units (ICUs) are intensely complicated environments characterized by a high degree of communication and collaboration between patient-care team members.1,2,3 Team members must often re-align and re-orient their knowledge about the patient, tasks, and goals to ensure high-quality care. Consequently, the need for timely communication of patient information often result in team members interrupting and being interrupted more frequently in the ICU.4 Information and communication technologies (ICT), such as, patient electronic medical records (EMR), pagers, smartphones create new opportunities for communication and dynamic updates for patient care; however, at the same time, they also introduce potential work interruptions through technology-mediated notifications (TMN) such as alarms, alerts, and pop-up messages.4,6,7 Gill and colleagues8 argue that a growth in mobile healthcare-related applications could lead to their increased use while communicating, collaborating, and disseminating tasks between healthcare providers. Although these technologies increase the ability to communicate and collaborate, they also introduce the possibility of someone being unnecessarily interrupted through a phone call or a text message.9 For instance, the introduction of wireless cellphones can lead to increased nurse-physician communication, but at the same time, cellphones can also interrupt the nurse’s or physician’s work. Further, the process involved with deciding whether to respond or to avoid the phone call can increase the already high cognitive load often experienced by ICU providers.10

Sasangohar and colleagues11 define interruptions as “externally or internally generated, unexpected events that may cause a break in the primary task, diverting attention to a related or unrelated secondary task, which can have both negative and positive effects on the interrupter’s or the interruptee’s main task.” (p. 3) Interruptions have also been called distractions, breaks-in-task, and disruptions.5,11,12,13 Several workflow-based studies in healthcare associate interruptions with negative impact on human cognition often leading to medical errors.15-16 For instance, 43% of medication errors are attributed to interruptions,16 and the Institute of Medicine (IOM) identifies interruptions as a potential factor contributing to medical errors.17 Consequently, existing research has focused on interruptions and factors that make them disruptive,5,11,12,13,14 often suggesting a set of design implications on when people can be interrupted.5,25,26,27,28 Although there is a predominantly negative view of interruptions caused by TMN, some studies acknowledge the important information a TMN may carry.9,13 Some examples of positive effects identified include contributing to increased safety, improvements in patient comfort, and increased accuracy.13 Because interruptions can be wanted or unwanted by a receiving ICU provider at any moment, we need to develop approaches for interruption management that increases the usefulness of the interruptions while reducing the negative effects on cognitive load.

Edwards and colleagues9 suggest taking advantage of properly designed asynchronous ICT that can mediate non-urgent interruptions, and divert a phone call to more opportune times. In this paper, we report on a study examining a potential asynchronous solution that involves presenting patient information in addition to the actual text message, to help the individual receiving the text to decide when to respond to a TMN. In the remaining portions of the paper,
we refer to the person performing an action leading to TMN as the “interrupter” or “sender”, and the person receiving the TMN as the “interrupted” or “receiver”. Our research investigation sought to investigate the following research question: “Would patient information such as vital signs in conjunction with a text message be useful in helping ICU providers decide whether or not to immediately interrupt their work to respond to a TMN?”

This paper is organized as follows: First, we review existing research on managing interruptions in the ICU. Second, we briefly describe our work on examining interruptions at an ICU site, including the identification of other additional patient information that is of interest to providers while responding to TMN. Third, we walk you through our study design methodology and results. Lastly, the paper discusses our findings and how our approach has the potential to help receivers better manage interruptions from TMN in the ICU.

Background

Researchers have extensively examined how people manage interruptions. For instance, Harr and Captelinin24 discuss the impact caused by the occurrence of an interruption beyond the interrupted activity through the creation of a “ripple effect”. This research suggests considering the social context in understanding peoples’ strategies for interruption management. On the other hand, some studies suggest considering the cognitive context as well while managing interruptions. Other studies argue that effective interruption management can be achieved through increased awareness among the colleagues.4,27 For instance, the aspect of “awareness related to action” can be used to reduce unwanted interruptions or facilitate wanted interruptions. The strategies people rely on to: (1) make their work visible to their colleagues, and (2) monitor what colleagues are doing, can help promote awareness, subsequently helping decide whether an interruption is wanted or unwanted at the moment.

Grandhi and Jones28 suggest considering “relational context” (such as, the content and urgency of interruption, interrupter-interrupted relationship and so on) while understanding and managing interruptions. The authors of this paper discuss the notion of how the receiver can associate a perceived cost and value to a TMN while deciding how to respond. Two interruption management paradigms are examined: (1) the impact reduction paradigm, which seeks to reduce the cost induced by an interruption by using techniques that either prevent or block TMN, and (2) the interruption value paradigm, which allows the receiver to evaluate an interruption based on its perceived cost and utility from a preview of useful information from the TMN.

Some recent innovations in ICT, such as, reminders and context-aware architecture, have been proposed to reduce interruptions through asynchronous communication. These systems, however, consider a limited number of factors from the receiver’s context. For instance, the communication service proposed by Dahl allows hospital workers to leave digital messages at relevant physical locations (such as patient bedside), so that the intended colleague can retrieve them at a later time. This approach is based on the assumption that hospital workers move between various locations mainly out of work priorities. However, other factors such as social activity at the location, content and urgency of each message, and the relationship of the timeliness of message reception to patient’s condition are not considered.

Previous research has focused primarily on identifying techniques for blocking interruptions in the ICU. For instance, researchers reported reduced interruptions with the introduction of an external signage or intervention that blocks face-face communication at selected locations, while others considered an awareness display that presents information to the interrupter about the workload currently experienced by the interruptee. These techniques, however, focus on the idea that both the sender and receiver are physically co-located and do not provide much assistance when they are geographically separated from each other. Further, these techniques require the receiver to remain within the same space as the interrupter when she does not want to be interrupted.

In this paper, we propose a solution that “adapts” the interruption value paradigm to include patient information for aiding “receiving” ICU providers, while improving their ability to manage interruptions from TMN with reduced cognitive load. This requires providing a preview of patient information (such as vital signs) along with the actual message to provide more context to the receiver while making
sense and deciding to interrupt an ongoing activity (Figure 1). In the next section, we briefly describe the field observations and follow-up focus group sessions, both of which were IRB approved and helped in the development of this approach.

**Examining Interruptions at an ICU Site**

**Observations:** We performed ethnographic field observation including contextual inquiry (of two ICU teams) in a 16-bed medical ICU used to monitor and treat critically ill patients. The observational notes were transcribed and coded using the Locales Framework. We viewed the situated nature of ICU work using the idea of dynamic social worlds. This helped us focus on the interaction at hand. We collaborated with two expert ICU physicians who helped validate our data interpretations and findings. We encourage readers to refer to previous work on understanding and designing for collaborative, sociotechnical environments.

The study site is comprised of electronic information sources such as EMR, patient charting and order notification system, bedside physiological monitors and infusion pumps. The existing technology systems supporting communication between the providers at the study site included electronic resources such as: (1) landline telephones (placed outside patient rooms, at nursing stations and conference rooms), (2) hospital-owned mobile ASCOM cellphones (which does not provide any information regarding the caller when a phone call is received), and (3) pagers (occasionally used for communication between the providers). The pagers are only used to convey callback phone numbers and notifications when a patient needs immediate attention (such as room number of a patient).

Below, we present an example involving a synchronous communication between a nurse and a resident regarding a patient. A nurse (N), who is currently at the nursing station, calls the patient’s resident (R) using cellphone. R, who is currently busy talking with his colleague regarding another critically ill patient, does not know for certain the urgency of the phone call compared to the ongoing face-face conversation. Since R cannot determine who is calling, he interrupts his face-face conversation and answers the call. We report an extract of this conversation.

N: Mr. S from 166 has positive blood culture for staph aureus.

R: What are his current vitals?

N: He has fever with temperature 101, blood pressure is 120 over 90, heart rate 65, respiratory rate is 11, and blood oxygen is 95% on ventilator.

R: All right, I will sign orders to start vancomycin and serial blood cultures.

N: Ok.

R had to interrupt his ongoing activity (face-face conversation with colleague) to respond to a phone call because he did not know who was calling him and the importance of the call. R was able to understand the condition of the patient and the urgency of the information conveyed based on additional patient information (such as current vital signs) provided to him by the nurse.

In an example of asynchronous communication – a resident who is participating in the morning rounds with his team receives a page message “0119865862”. The resident understands that the message is a callback number. Since he does not know the actual patient information that will be conveyed in the phone call, and since he does not know how urgent it is to call back, he steps away from the team to call the number. This results in the resident interrupting his current activity and missing some of the morning rounds discussion.

If the resident in the above example had received a visual cue containing the lab result and vital signs from the nurse before initiating a phone call, this would have been less disruptive. These cues will improve the contextual awareness and also provide implicit information on the priority with which the resident had to respond. For instance, seeing an abnormal temperature, heart rate, blood oxygen saturation in addition to the lab results can indicate to the resident that the patient is having an inflammatory sepsis response. That is, the extra patient-centric information such as vital signs that were discussed during the phone call could have been provided beforehand along with the actual message to expedite the resident’s understanding and decision-making of next steps. Further, the resident would have had more freedom in planning and strategizing when to interrupt an ongoing activity while responding.

**Focus Group:** We conducted a focus group session (FG1) with five ICU providers to understand the type of patient information that receiving providers are interested in while deciding how they responded to TMN. This was
followed by another focus group session (FG2) with two ICU providers and five user experience designers to create a design template that will be presented as preview text message notifications (Figure 2), including fictitious scenarios and information content for the sample text messages used to evaluate our designed solution.

In general, the information shared between providers include: lab/radiology results, patient’s response to medication, procedure completion and corresponding patient’s condition, and follow-up/consult with a specialist. One of the findings from FG1 also pointed to other additional details that can persuade the receiver to pay more attention while deciding to respond to a TMN. For instance, the information perceived by the sender based on the patient’s current condition, i.e., the indirect information perceived from the patient is often used to indicate the “priority/urgency of the message” during communication. FG1 also generated a set of sample text messages that can be shared between ICU providers. Several benefits and challenges to multiple versions of design templates were discussed and clarified in FG2. The design considerations from FG2 included (Figure 2): (1) explicitly depicting the importance of a message with respect to patient’s current condition by coloring the background of the preview notification (red: high; yellow: low; grey: none), (2) explicit color-coding of content to direct the attention of provider to important information that is abnormal or needs attention (red), (3) details on the patient and the sender of the message, and (4) placement of the patient’s current vital signs and the contents of the actual message. The focus group sessions contributed to the generation of fictitious scenarios and sample messages that can be conveyed as a text message notification.

Figure 2. Patient-enhanced notification depicting the actual message (+BC staph aureus) along with additional information such as vital signs and priority of the message (red background color). Content is explicitly color-coded to direct attention to parts of the actual message (+BC) and vital signs (101).

Based on our observations and focus groups, we propose a solution that involves presenting additional information obtained directly (from EMR) and indirectly (such as nurse’s perception) from the patient, along with the actual message as a “preview text message” to improve the receiver’s understanding and awareness of the patient, eventually contributing to making rapid decisions on when to respond to text message notifications. We refer to this “preview text message” notification with additional information as “patient-enhanced notifications.” The remainder of the paper describes the study design and our findings from an initial evaluation of the patient-enhanced notifications.

Study Design

Participants: Twelve ICU providers from the same study site as our fieldwork and focus group sessions participated in this study. The participants of the study are not the same as those who participated in the focus group sessions. The participant included: residents (P1, P2), nurse practitioners (P3, P4), fellows (P5, P6), registered nurses (P7–11), and staff nurse (P12).

Study Procedure: The study consisted of two activities: an evaluation of the patient-enhanced notification and an interview.

Evaluation session: This session was conducted in a quiet conference room where participants were required to use a web-browser (on a workstation) to evaluate the design of patient notifications. The evaluation required each
participant to view a fictitious scenario along with an image of a patient notification (Figure 3). Participants did not receive any instruction manual explaining the design of the patient notifications.

Participants responded to a set of multiple-choice questions. The multiple-choice questions included: (1) How would you react to this notification message? The response choices for this question included immediately, respond later, or do nothing, (2) When do you prefer receiving this notification? The response options for the questions were chosen based on existing research\(^9\)\(^,\)^\(^27\)\(^,\)^\(^41\), where, participants were asked to indicate whether a particular message in a given situation should be delivered immediately or at a future scheduled time. This session included a within-subjects design that required all the participants to respond to all the multiple-choice questions based on all the fictitious scenarios presented along with images of all text message notifications (Figure 3). All participants played the role of receiver of text messages. All the participants also rated their perceived urgency of each message on a scale from 1–10.

![Figure 3. Screenshot of what the participants saw while completing the evaluation session](image)

We utilized two scenarios that required the participants to imagine a receiver’s situation while receiving a text message: (1) high cognitive load and present in patient room and, (2) normal cognitive load and in cafeteria. For instance, a provider playing the role of a receiver was presented a scenario and image as in Figure 3. The participant was then asked to report on when he/she will respond (immediately vs. later vs. do nothing). We presented the image of only one message at a time to examine the initial reaction of providers to the additional information in the patient-enhanced notifications (tested through the study conditions described below).

**Study conditions:** We varied the additional information that can be shared along with the actual text message to assess their value of particular type of additional information. We varied the order that the study conditions were presented between participants to avoid order effects. This study focused on providing the following four levels of additional information in text messages:

1. **None:** In this condition, providers saw only the message with no additional information. The background of the text message was colored grey.

2. **Priority:** This condition provided additional information such as the priority associated with the message. The levels of this condition included: Low and High. Explicit color-coding was used to depict priority information. For instance, the background of text messages was colored red or yellow to depict urgency of the message as high or low respectively.

3. **Vitals:** This condition provided additional information on patient vital signs such as heart rate, blood pressure, temperature, respiratory rate, and blood oxygen content. The levels in this condition included: With and Without. The patient vital signs were again explicitly color-coded with red color to indicate abnormal values.

4. **Combination of priority and vitals:** This condition included a combination of conditions 2 and 3 (Low vs. High and With vs. Without). All the participants were provided with the same study conditions, in a randomized order.
Data collection and analysis: Participants responded to questions administered through the TypeForm website. The TypeForm website has a backend database tool for collecting participant responses. Quantitative data was analyzed using SPSS v.21. Participant responses to the multiple-choice questions were coded for performing descriptive and inferential analysis (1 = respond immediately, 2 = respond later, 3 = do nothing). We explored the receiver’s willingness to respond immediately to a message with different types of additional information. Friedman’s test with the additional information such as: (1) priority (low vs. high), (2) inclusion of vital signs (with vs. without), (3) combination of both (Priority x Vitals), (4) perceived urgency of message (1 – 10), (5) the 2 scenarios, (6) designation of the participant (doctor vs. nurse), and (7) message viewing order (1 – 10) was used to examine the receiver’s preferences for responding to text messages (immediately vs. later vs. do nothing). The rated urgency of these messages tended to be of low variance across the raters and did not cause confusion (Figure 4).

Interview session: The interview session occurred at the end of every evaluation session. In this session, all the participants responded to demographics-related questions and open-ended debriefing interview questions. The open-ended questions asked about: (1) the overall experience interacting with the designed text message notifications, (2) benefits of additional information while understanding and improving patient-situation awareness, and (3) how the additional information provided along with the actual text message aids in deciding when and how to respond to a text message alert or notification.

Data collection and analysis: Participant responses to the interview questions were audio-recorded and later transcribed for analysis. Qualitative data was coded and analyzed using ATLAS.ti. The post-evaluation interview session provided more information on how the additional information provided along with the actual message aid in deciding when and how to respond to a text message alert or notification. We report themes emerging from the coding of participants responses. Participant responses were also coded to identify suggestions that we used to help us provide some design suggestions for the patient notifications (Table 1).

Results

Quantitative

There was no significant difference among providers in perceptions of how important it is to respond immediately while responding to messages with and without additional priority information ($Z = -1.630, p = .103$). Specific to messages that had additional priority information provided as high or low, providers preferred responding to text messages immediately ($Z = -3.503, p = .04$) if the priority of the message was depicted as high ($M_{High} = 67.19\%$) as compared with those depicted as low ($M_{Low} = 32.81\%$). No significant difference in response preferences was noticed based on the scenarios, or the designation of the participant (doctor vs. nurse). Providers preferred responding to messages immediately ($Z = -3.645, p = .037$) if vital signs were included ($M = 60\%$) as opposed to not including the vital signs ($M = 40\%$) with the actual message. A more detailed exploration of the interaction between priority (high vs. low) and patient vital signs (with vs. without) demonstrates that the presence of patient vital signs coupled with priority of the message affected provider’s preferences in responding immediately to text messages ($\chi^2 (3) = 9.510, p = .023$; Figure 5).

That is, providers preferred interrupting their ongoing activity to respond to text messages based on the priority explicitly conveyed through the background color, and implicitly perceived from the patient vital signs. For instance, receivers interrupted themselves less often for low priority as opposed to high priority messages when they knew the patient vital signs. On the contrary, they had similar preferences for immediate interruption for messages with no vital signs irrespective of the priority of the message. Overall, their preference was also backed by their perception of the urgency with which they had to respond based on the contents of the actual message in relation to vital signs.

Qualitative results

Color-coding: Providers found it easy to rapidly understand and perceive the information conveyed through colors in patient notifications. For instance, P5 noted, “if I see something colored red, I automatically think it is abnormal. I was able to quickly understand that a vital sign is abnormal. Although I don’t really need the coloring, I know what is normal and abnormal. I think the color will help me improve my speed with which I perceive the information presented. This can be very helpful when all I have to do is glance at the phone while I am in the middle of something.”

Using patient vital signs: Providers reported patient vital signs as important information that they can use to rapidly understand a patient’s condition from the message. For instance, P2 noted, “the first thing I want to know when a patient’s lab result is abnormal is his current vitals. Instead of trying to look into EMR or call the nurse, it will greatly save me time and effort if I can just glance at my device and see the message and vitals together. This will help me to quickly determine what has to be done next, do I respond to the message at the moment or continue with whatever I was doing.”

Providers reported glancing at patient vital signs to rapidly determine the time they had to respond to the message. For instance, P1 noted, “if I am in the middle of rounds and if I receive a message with say a lab result and current vital signs, then I will be able to quickly understand the patient’s current condition and how it relates to the lab result, what has to be done and when. If it is something important and needs immediate attention, then I can step away from rounds, otherwise I can respond later to the message. However, I am concerned if I will remember to respond later, you know what I mean?” (this has been addressed as a design suggestion in Table 1)

Explicitly stating information on the urgency of a message shared: Providers reported not paying attention to the background color of patient notifications used to explicitly indicate the priority of the message. For instance, P1 noted, “the patient vital signs are enough for me to know how important the message is. Besides, I might think a message is not really important while a nurse might consider the same message as really important. This is probably why I never noticed the background color of the messages. This might be useful if I see a series of text messages all together. That way I can differentiate which ones are important and which ones are less important.”

Overall, participants reported positive preference for receiving a text message with color-coded patient vital signs during the interview session. In addition, participants provided suggestions, which was used to derive design suggestions for improving the evaluated design of the patient-enhanced notifications (Table 1).

Discussion

There are multiple approaches to addressing interruptions. Some researchers have adopted the impact reduction paradigm to manage interruptions, while others have adopted the interruption value paradigm to aid in the decision-making of response to the TMNs. While using the former can have an impact on the quality of patient care owing to the complex temporal and mobile nature of ICU work, the latter can add to the cognitive load already experienced by the receiving providers while evaluating the cost and utility of every TMN. We chose to adopt the interruption value paradigm because of the freedom it allows for a receiving provider to determine how to respond to a TMN. By using a combination of fieldwork and focus group sessions, we proposed patient-enhanced notifications as a tool to circumvent the added cognitive load while managing interruptions from TMNs. This is a novel contribution that supports offloading the cognitive expense (such as memory, inferential strategies, and knowledge) faced by ICU providers when using technology-mediated communication. Further, our work contributes to the Institute of Medicine’s (IOM) requirements of introducing health IT that does not increase cognitive load or alarm fatigue in the ICU.
Table 1. Design suggestions

<table>
<thead>
<tr>
<th>Description</th>
<th>Design Suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurses suggested that the information that is often shared between one</td>
<td>Display a previous intervention result in addition to current result to provide</td>
</tr>
<tr>
<td>another is related to repeated interventions, such as, labs. P8 noted,</td>
<td>clearer understanding of the patient’s progress over time.</td>
</tr>
<tr>
<td>“Sometimes I prefer comparing a previous lab result just to see how the</td>
<td></td>
</tr>
<tr>
<td>patient has been doing over time. Possibly it will be helpful if I could</td>
<td></td>
</tr>
<tr>
<td>see just the previous and current lab result. I would anyways do this by</td>
<td></td>
</tr>
<tr>
<td>going to the EMR if the results don’t look good.”</td>
<td></td>
</tr>
<tr>
<td>Nurses found the term “Dismiss” unclear. They wanted an option that will</td>
<td>Replace the response option “Dismiss” to “Acknowledge”</td>
</tr>
<tr>
<td>allow a sender know that the receiver has seen his/her message. P5 noted,</td>
<td></td>
</tr>
<tr>
<td>“let’s say the resident is busy and can’t respond to my message immediately. If I can see that the resident has acknowledged my message, then I can move on and work on other activities since I have completed my part of work.”</td>
<td></td>
</tr>
<tr>
<td>All providers reported on the value of sending and receiving a response to</td>
<td>Dynamically change the response options to “yes” and “no, respond” if the actual message is a question.</td>
</tr>
<tr>
<td>a specific question from the colleague. However, “Dismiss” and “Respond”</td>
<td></td>
</tr>
<tr>
<td>buttons were found not to support this functionality. P2 noted, “it will</td>
<td></td>
</tr>
<tr>
<td>make more sense to me if I see a question in the message and I respond with a yes or no, maybe for the no option, I can explain why I chose that response to the recipient.”</td>
<td></td>
</tr>
<tr>
<td>Nurses raised the issue of difference between communicating multiple topics</td>
<td>Change “Respond” option to allow providers to choose from a set of options: Callback and Text.</td>
</tr>
<tr>
<td>within a phone conversation as opposed to conveying a single topic through</td>
<td></td>
</tr>
<tr>
<td>a text message.</td>
<td></td>
</tr>
<tr>
<td>Residents and nurses raised the issue of trying to remember to respond to</td>
<td>Add an option, where providers can set a time for the patient notification to pop up again; such as, “Remind me in 1 hr.”</td>
</tr>
<tr>
<td>a patient notification if they chose “Acknowledge” at any point of time.</td>
<td></td>
</tr>
</tbody>
</table>

Our evaluation study demonstrated that ICU providers responded immediately to messages that provided more contextual information such as patient vital signs in addition to the actual message. The urgency as portrayed using color was intended to provide an indication for the providers to “glance at” and determine the importance of the notification. However, providers reported that explicitly depicting urgency might not be helpful initially, if only a single message is seen on the mobile device – suggesting that such a feature will be learnt over time. Providers reported finding it easy to direct their attention to abnormal values since they were colored red – similar to their mental model of attributing red color with abnormality. Further, providers reported being able to perceive the urgency of a message implicitly based on the patient vital signs and the actual message. Providers highlighted the inclusion of the patient vital signs with a text message as a significant contributor in aiding their decision on how to respond to the message – patient vital signs were considered as important information that can be “glanced at” during the decision-making of any patient activity. Based on our findings, we also provide some design suggestion for the patient-enhanced notifications (Table 1). Most of these suggestions revolved around the idea of providing more contexts to the receiver as well as simplifying the response process.

**Study Limitations:** This study chose a subset of additional patient information, cognitive, and environmental situations to present. For example, we chose not to vary the time of the day or relationship between the interrupter and patient. We did this to gain more experimental control. Also, the specific options we chose can be implemented in the ICU using existing technologies. For instance, the patient initials, room number, and vital signs can be automatically populated if the device is paired with the EMR; leaving the sender to only type in the actual message and choose the importance of the message with respect to his/her understanding of the patient’s current condition. Another limitation is that this was laboratory-based. Consequently, how providers respond in a real-world setting can be affected by a variety of factors that we may be unaware of.
Conclusion
Managing interruptions is a major challenge in a busy, information-intensive ICU environment. In this paper, we adapt the interruption value paradigm\textsuperscript{27} to suggest an asynchronous technology solution that can help receivers better manage interruptions. We were interested in understanding what features would best support a receiver’s decision-making process in term of deciding whether or not to immediately respond to a TMN. Our proposed solution, also known as patient-enhanced notifications, used additional patient information (current vital signs and priority/urgency of the message perceived by interrupter based on patient’s current condition) provided via text message previews. Our study showed that providers were receptive to the additional information and that it helped them in deciding whether or not to immediately respond to a TMN. We also highlighted some design suggestions, one of which points to combining asynchronous and synchronous communication capabilities of a technology tool. In summary, we believe that introducing an asynchronous mechanism that prompts a response phone call, by sending patient-enhanced notifications, has two positive outcomes. First, it can support a conversation of multiple topics and second, it can support the interrupted in better managing the interruptions generated by the TMN. As to future work, we intend to further examine the differences in better managing interruptions between an asynchronous solution (that facilitates conversation based on only one topic) as opposed to a synchronous solution (that facilitates conversation of multiple topics).

Acknowledgments
We thank the ICU providers for allowing us to observe them. We also thank the ICU providers for evaluating the design of patient notifications and responding to interview questions.

References
Predicting Negative Events: Using Post-discharge Data to Detect High-Risk Patients

Lina Sulieman, MS¹, Daniel Fabbri, PhD¹, Fei Wang, PhD², Jianying Hu, PhD³, Bradley A. Malin, PhD¹
¹ Vanderbilt University, Nashville, TN; ² Cornell University, New York, NY; ³ IBM T. J. Watson Research Center, New York, NY

Abstract

Predicting negative outcomes, such as readmission or death, and detecting high-risk patients are important yet challenging problems in medical informatics. Various models have been proposed to detect high-risk patients; however, the state of the art relies on patient information collected before or at the time of discharge to predict future outcomes. In this paper, we investigate the effect of including data generated post discharge to predict negative outcomes. Specifically, we focus on two types of patients admitted to the Vanderbilt University Medical Center between 2010-2013: i) those with an acute event - 704 hip fractures and ii) those with chronic problems – 5250 congestive heart failure (CHF) patients. We show that the post-discharge model improved the AUC of the LACE index, a standard readmission scoring function, by 20 - 30%. Moreover, the new model resulted in higher AUCs by 15 - 27% for hip fracture and 10 - 12% for CHF compared to standard models.

Introduction

Predicting events associated with negative outcomes, such as readmission or death, post discharge is challenging. Traditionally, negative outcome prediction systems are executed at the time of discharge to identify high-risk patients¹. However, such systems are limited in their applicability because patient status often changes after their discharge and risk prediction models are not amended to incorporate such information. This lack of knowledge can result in risk assessment errors and potential readmission penalties, under Meaningful Use regulation, calculated via a payment adjustment factor².

Care providers, as well as the administrators of healthcare organizations, are keen on determining which patients might experience complications that lead to readmission or death, so that they may allocate additional resources to the patients and intervene before a negative outcome transpires³. Unfortunately, resources are limited and decision makers within health organizations (e.g. physicians and care coordinators) specify a subset high-risk patients who can receive special attention while other high-risk patients do not have access to such resources⁴,⁵. Accurately identifying, as well as prioritizing, which patients should be assigned assistance is an important informatics problem.

To investigate the effect of post discharge data in predicting negative outcomes, we focus on several core questions. First, what post-discharge information is available for prediction? Second, how can a post-discharge model be formulated? And, finally, how do different time periods of post-discharge information impact such predictions? In doing so, we examine which post-discharge features are the most important drivers of negative outcome predictions.

We evaluated standard and post-discharge prediction models using three years of data from Vanderbilt University Medical Center’s (VUMC’s) electronic medical record (EMR) system for two phenotypes: 1) an acute condition in the form of a hip fracture and 2) a chronic progressive disorder in the form of congestive heart failure (CHF). The results demonstrate that:

- Running the prediction at successive post-discharge days, and including post-discharge clinical information in the prediction model, outperforms state-of-the-art “at discharge” models, such as LACE⁶,⁷;
- The importance of post-discharge clinical features grows as the prediction horizon for negative events is pushed further into the future; and
- Higher utilization of clinical resources (e.g. appointments and medications) are correlated with a negative outcome.

Background

The number of proposed risk prediction models has increased dramatically over the last decade. Kansagara and colleagues performed a comprehensive systemic review to evaluate the performance of risk prediction models and their suitability for clinical use¹. To evaluate the performance of the risk model, studies usually compare their models to an established model such as LACE. This is a readmission index that provides a risk score to predict the
readmission or death, specifically for CHF patients, using length of stay (L), the acuity of admission (A), comorbidity score (C), and the number of the emergency department visits (E)\textsuperscript{6,7}.

Several studies have shown that post-discharge data can assist in the prediction of negative events in special circumstances. For instance, certain studies focused on surgical quality assessment at the point of discharge observed that over a quarter of the complications are diagnosed post-discharge\textsuperscript{8-10}. In another study, it was found that post-discharge data could improve the prediction of the presence, as well as the severity, of spasticity in upper limbs in the year following a stroke\textsuperscript{11}.

While post-discharge data has rarely been used in readmission prediction, Hersh and colleagues performed a systemic review about the post-discharge environment and its relation to readmission after heart failure\textsuperscript{12}. They reviewed 26 studies published between 1985 and 2011 to evaluate the importance of integrating post-discharge environment in the heart failure readmission model. In the review, only 7 studies included post-discharge data and focused mainly on whether the patient had a primary care provider. They concluded that the socio-economics of the post-discharge environment is a key indicator that affects readmission probability. Another factor that has been found to correlate with readmission is follow-up after discharge. Specifically, patients with a larger number of early follow-ups tend to have a lower likelihood of unplanned readmission, especially for patients with a greater collection of comorbidities\textsuperscript{13-16}.

Methods

In this section, we describe the risk prediction models. First, we describe methods to construct features using data from different post-discharge time windows. Second, we use these features to predict negative events occurring up to a prediction window.

Prediction Model Composition

We represent longitudinal EMR data with an $M^{NT}$ matrix, where $T$ is the length of time for extracted EMR data and $N$ is the number of patients in the cohort. Each row in the matrix represents one patient vector denoted by $t$. For each patient $p$, we divide the longitudinal medical record vector $t$ into three temporal bins: before admission, during admission, and after discharge. For each model, we define an observation window (OW) that specifies the time from which to extract features. Using combination of features, we predict whether the patients would have a negative outcome within a prediction window (PW), where the window starts at the discharge day and ends at the prediction point (PP).

For the purposes of this investigation, we represent outcomes as a dichotomous variable. A patient has a negative outcome if he/she experiences a negative event (-1) in the prediction window and has a non-negative outcome (+1) otherwise. Figure 1 visualizes model settings in which both patients experience negative outcomes, but only the patient in Figure 1b has a negative outcome in the prediction window.

![Figure 1: Prediction model – prediction point before negative outcome](image1)

**Model 1: LACE:** Starting with the most common method in the literature, we build a prediction model using LACE, where $E$ was restricted to the number of the emergency department visits in the past 6 months. LACE assigns points to each variable based on its value, and calculates probabilities by using regression models\textsuperscript{7}. Patients with a score greater than 10 are considered to be high risk. We retrieved the features utilized by LACE from the EMR and calculated the risk score, which was fed as a feature into the prediction model. We consider LACE as the baseline for performance comparison.
Model 2: Before-Admission Model (BAM): To learn whether prior health status can be applied to forecast the future health status of patients, we use only data from before the admission to predict the outcome. Each patient’s entry is assigned to a pre-admission feature vector denoted by \( b \). The BAM matrix is visualized in Figure 2a.

Model 3: During-Admission Model (DAM): We investigated whether the data collected about investigated phenotype is sufficient to predict the outcome. The observation window begins at admission and ends at discharge and is visualized in Figure 2b.

Model 4: At-Discharge Model (ADM): This model incorporates data from before and during admission to predict the outcome, and is visualized in Figure 2c.

Model 5: Before, During admission, and After discharge Model (BDAM): The last model incorporates post-discharge data to predict the patient's outcome, which is visualized in Figure 2d. There are several key parameters to this model: 1) the time window post-discharge to include data, which we refer to as a checkpoint, 2) the length of the prediction window, and 3) ensuring that data prior to negative events does not introduce biased knowledge about an upcoming negative event. Regarding point (3), we remove data from days immediately before the end of the window, which we refer to as the gap. Thus, the model can be specified with a checkpoint \( C \), gap \( G \), a prediction point \( PP \) and prediction window \( PW \).

**Overview of Features**

For the patient records in this study, we extracted the demographics (age, gender) and data ranging from one year before to one year after the first documented incidence of the phenotype under investigation (i.e., hip fracture and CHF). For each patient, we extracted the number of resources that were allocated for treatment, including: 1) medications, 2) laboratory tests, 3) appointments, 4) previous admissions, 5) the average of previous length of stays (LOS), 6) days since the last admission, and 7) the count of the International Classification of Diseases, Ninth Revision (ICD-9)\(^{17} \) in each of the 20 chapters. Table 1 summarizes the features and the temporal period to which they correspond (e.g., before admission, during admission, or after discharge). In addition, we extracted the number of documents, grouped by their type, that were created and stored in the EMR during hospitalization.

Table 2 lists the extracted lab tests, their normal ranges, and the diagnostic purpose of the test. We retrieved the average values of the most common lab tests that were ordered for 80% of the patients during admission and after discharge, including carbon dioxide levels (CO\(_2\)), creatinine, glucose, hematocrit or packed cell volume (PCV), partial thromboplastin time (PTT), potassium (K), sodium (Na). These lab tests are ordered by clinicians to evaluate heart and kidney functionality, electrolyte balances, and blood clotting timing.

1171
Table 1: Summary of the features included in the models with the observation window taken from. The symbols *, +, and - represents extracted from before, during and after bins respectively.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Values</th>
<th>Feature Bin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Age and gender.</td>
<td>*</td>
</tr>
<tr>
<td>Laboratory tests</td>
<td>Number of laboratory tests.</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Average values of: glucose, creatinine, partial thromboplastin time (PTT), hematocrit or packed cell volume (PCV), Carbon Dioxide levels (CO₂), potassium (K), and sodium (Na).</td>
<td>-</td>
</tr>
<tr>
<td>Medication</td>
<td>Number of medications prescribed for the patient.</td>
<td>*</td>
</tr>
<tr>
<td>ICD</td>
<td>The count of ICD9 in each of 20 chapters.</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>ICD deviation post-fracture (the ratio of ICD chapters number in an appointment after discharge to the average number of ICD chapters before hip fracture incidence).</td>
<td>-</td>
</tr>
<tr>
<td>Routine care</td>
<td>The average of Braden score, the number of ECG tests, the number of times a patient received respiratory care.</td>
<td>+</td>
</tr>
<tr>
<td>Admission</td>
<td>Length of Stay (LOS).</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Last day of previous admission.</td>
<td>*</td>
</tr>
<tr>
<td>Appointment</td>
<td>The number of appointments.</td>
<td>*</td>
</tr>
<tr>
<td>Documents and</td>
<td>The number of communication message</td>
<td>-</td>
</tr>
<tr>
<td>communication</td>
<td>The number of documents initiated for per document type.</td>
<td>+</td>
</tr>
<tr>
<td>Post-discharge time</td>
<td>Number of days since discharge.</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Number of days until prediction point</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Common lab tests, their normal values, and the diagnostic purpose.

<table>
<thead>
<tr>
<th>Lab Test Name</th>
<th>Normal values</th>
<th>Purpose</th>
<th>Abnormal values reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creatinine</td>
<td>Male: 1.3 mg/dL, Female: 1.1 mg/dL</td>
<td>Test kidneys functionality</td>
<td>Higher than normal level is an indicator of kidney malfunction such as kidney failure, blocked urinary tract, and kidney damage.</td>
</tr>
<tr>
<td>Partial thromboplastin time (PTT)</td>
<td>25-35 seconds</td>
<td>Measuring the time that the blood takes to clot</td>
<td>Abnormal or long PTT time indicate bleeding disorder or disorder in clotting process</td>
</tr>
<tr>
<td>Hematocrit or packed cell volume (PCV)</td>
<td>Male: 55%, Female: 42%</td>
<td>Measuring the percentage of Red Blood Cells (RBC) in blood</td>
<td>Low PCV: indicator of anemia, over-hydration, and destruction of RBC High PCV indicator of dehydration, congenital heart disease, or abnormal increase in RBC</td>
</tr>
<tr>
<td>Carbon Dioxide levels (CO₂)</td>
<td>23 to 29 mEq/L</td>
<td>Detecting the body’s electrolytes imbalance</td>
<td>Low levels: indicator of acidosis, Kidney disease High level: indicator of breathing disorders, hyperaldosteronism</td>
</tr>
<tr>
<td>Potassium (K)</td>
<td>3.7 to 5.2 mEq/L</td>
<td>Assessing the kidney and heart functions.</td>
<td>Low levels: Chronic diarrhea, renal artery stenosis, diuretics High levels: blood transfusion, Kidney failure, acidosis</td>
</tr>
<tr>
<td>Sodium (Na)</td>
<td>135 -145 mEq/L</td>
<td>Measuring balance between sodium and water in consumed foods and drinks</td>
<td>Hyponatremia (Na &lt; 135 mEq/L): kidney disease, heart failure, or ketones in blood from starvation Hypernatremia (Na &gt; 145 mEq/L): dehydration, severe vomiting, or diabetes</td>
</tr>
</tbody>
</table>

Negative Outcome Prediction Over Time

To study the performance of the models over time (specifically the post-discharge data), we predicted the outcome for different prediction points and prediction windows ranging from seven days to one year.

Model Implementation

In this section, we describe the data extraction and provided high-level overview of the prediction algorithms.

Data extraction

We extracted patients who were diagnosed with a hip fracture or CHF, using the 820.* and 428.* range of ICD9 codes, respectively. We excluded patients who had no encounters before admission and after discharge to triage patients who visited the medical center for only that admission. We also excluded repeated admissions for these phenotypes with the aim to only analyze the first admission.

Input matrix: We construct the BAM, DAM, ADM, and BDAM input matrices. For the BDAM matrix, different checkpoint days can influence the prediction’s performance. To remedy this issue, we uniformly sampled checkpoints at random between the discharge and prediction points and averaged the results. This random day
sampling approach provides a viable option for evaluating the model even though the model changes as different amounts of post-discharge data are included.

**Changing the prediction window length:** To understand the change in risk and the effects of the phenotype over time, we varied the prediction window length from 7 to 365 days.

**Model implementation:** We used a random forest classifier from Scikit-learn and conducted a 5-fold cross-validation. For random forest parameters, we used 500 trees with 15 maximum depth and 15 as the minimum split.

To construct the BDAM matrix, we enforced a gap of 5 days for all prediction points except seven days, for which we used a gap of three because it yields four days from which to sample compared to sampling from two days. The values in the post-discharge data depend on the checkpoint location. As shown in Figure 3, different checkpoint locations lead to different BDAM entries for the same dataset (Figures 3a and 3b exhibit the vectors constructed with different checkpoints). To minimize the effect of randomness, we built 100 matrices at each prediction point, randomly sampled checkpoints for each matrix and averaged the area under the receiver operator characteristic curves (AUCs).

**Figure 3: Building BDAM vector using two checkpoints (a) an example of a randomly sampled checkpoint and (b) an alternative sampled checkpoint**

**Important features and their relationship to outcomes**

To characterize the importance of the features in the post-discharge model, we extracted the features that, on average, have the highest importance across all folds. We analyzed the importance of the features that comprise around 50% of the total feature importance. Additionally, we analyzed the non-linearity of features with partial dependence plots. The partial dependence shows the dependence of prediction approximation on a subset of the input variables. It finds the marginal average of the prediction to identify the effect of chosen subsets of features on the prediction probability after accounting for the rest of the input features. For a given predictor, the y values in the partial plot show the average of prediction probability across all trees in the forest.

**Results**

We begin by summarizing the patient populations and their negative outcome rates. Then, we present the performance of the models, with LACE as the baseline. Finally, we report on the importance of the features incorporated in the model and the outcome.

**Patient Population**

This study was based on patients who were admitted to VUMC between 2010 and 2013. We included patients aged 65 years and older and were diagnosed with either a hip fracture or CHF. We excluded patients who did not have an admission on the onset date of the phenotype. This selection criteria yielded 704 hip fracture patients and 5250 CHF patients.

Around 25%, and 21% of hip fracture and CHF patients, respectively, had a negative outcome within 90 days after discharge. As shown in Figure 4, the one-year survival rates exhibited a similar trend. For patients who exhibited a negative outcome within the first 7 days, more than half were admitted to the emergency department or died within the first three days after discharge.
Before Discharge Model Results

Figure 5 reports on the AUC for predicting the negative outcome within one year using LACE, before admission, during admission, and at discharge) models for hip fracture and CHF patients. The x-axis shows the prediction points at which the model was run, while the y-axis corresponds to the average AUC. Across all prediction points, LACE exhibited the lowest AUC values for both cohorts. The before model has similar AUC values as LACE for hip fracture patients, while it has higher AUC than LACE when applied on the CHF cohort. However, the during and at discharge both outperform the LACE and before models. The during model exhibits almost the same AUC as the at-discharge model when predicting the outcome for the CHF patients. By contrast, at discharge model has a higher AUC than the during model for the hip fracture patients.

In both cohorts, the before model had the lowest AUC within the first 20 days. After 20 days, the AUC values increased slightly. The during and at-discharge models exhibited the same trend in performance over time. In CHF, the during model and at-discharge models had the highest AUC during the first two weeks of prediction. Afterwards, the AUC values decline slightly and smoothly until they reach their lowest point at the one-year prediction. By contrast, the AUC values of the during model and at-discharge were the lowest within the first two weeks for the hip fracture patients and increased slightly until they reached their highest AUC values at the one-year prediction.

Post-Discharge Model Results

Next, we analyzed the post-discharge model in two ways: (i) using a single feature representing the days since discharge (post-discharge time), and (ii) all post-discharge data including days since discharge. We analyzed the
results in these two phases given the importance of time in the post-discharge model (i.e., the longer you are out of the hospital, the less likely you are to return).

Figure 6 depicts the AUC values for specific checkpoints for the 7, 14, 22, and 30 days prediction points. The x-axis corresponds to the selected checkpoint. The red dashed-line represents the average AUC for post-discharge model at a given checkpoint, while the diamond and triangles correspond to the AUC of at-discharge model and the traditional LACE model, respectively. The blue solid line represents the AUC for the post-discharge time only model. It can be seen that the BDAM model outperformed all models. Moreover, the AUC values increased as additional post-discharge data are included.

Figure 6: AUC values for BDAM model applied at different successive checkpoints for hip fracture patients at 7, 14, 22 and 30 days negative outcome.

Figure 7 depicts the performance of the LACE, ADM (at discharge) and BDAM (post-discharge) models for different prediction points (where the post-discharge data are averaged across the checkpoints). The upper grey line, lower grey line, and solid black line represent the minimum, maximum, and average AUC, respectively. Both the BDAM and the ADM models performed better than the LACE model by 20 - 30% AUC. In this setting, BDAM had a higher AUC than ADM at all prediction points by 15.8 - 26.5% for hip fracture and 9.7 - 12.1% for CHF patients. Using BDAM, predicting the negative outcome within 7 days had a higher AUC compared to predictions within 30 days.

BDAM results trended differently for the two phenotypes. For hip fracture, the AUC decreased until it reached the lowest values at 21 days, then increased between 21 and 60 days, staying relatively constant afterwards. For CHF, the AUC decreased until 21 days.

Feature Importance

For each prediction point, we retrieved the features that, on average, exhibited the highest importance. At all prediction points in both cohorts, days from discharge and days until prediction point displayed the highest importance. The number of appointments after discharge was the third most important predictor for a negative event within 30 days while clinical communication was the forth or fifth important feature within the same time range.

In hip fracture patients, age is one of the top 10 predictors of a negative outcome within 30 days. Diagnosing hip fracture patients with infectious, blood stream, genitourinary, or circulatory diseases during admission were among the top predictors for readmission within 30 days. In CHF patients, the average number of labs during admission (e.g., creatinine, K, Na, and CO₂) were strong predictors for negative events within 60 days.

Figure 8 depicts the partial dependency plots for the most important features for 7, 21, 30, and 60 days for hip fracture patients. In the partial dependence plot, the x-axis shows the values of the variable. For a given x value, the
y value indicates the probability of the negative outcome after accounting for the values of the other input variables. A small y value indicates a low probability of positive outcome, while a large y indicates a high probability of a positive outcome.

For example, patients who scheduled a small number of appointments after discharge had a higher probability of experiencing a negative outcome. A low and high number of appointments scheduled before discharge were associated with a high probability of negative outcome. In addition, a low and high quantity of post-discharge communications had a high probability of a negative outcome. During admission, patients who were prescribed a larger number of medications and lab tests, in comparison to other patients in the cohort, had a higher probability of having a negative outcome. The partial dependency plots of lab values such as K, PVC, and creatinine depict that abnormal values were associated with a high probability of a negative outcome.

Discussion

Predicting negative outcomes and detecting high-risk patients are challenging problems. Our findings demonstrate that including more information about a patient’s post-discharge status may increase the performance of such predictions. This finding is further supported by the observation that the LACE model, while simple to implement, is likely to neglect many key features that can enhance predictions.

One of the critical discoveries made in this study is that the post-discharge time is a strong predictor on its own for a negative outcome. This result affirms the observation that the longer a patient remains out of the hospital, the less likely they will be readmitted at a future point in time. However, time from discharge in isolation lacks important clinical information, which if included, can further improve prediction quality. Moreover, time until prediction quantifies the amount of information collected post-discharge that is not included in time from discharge.

One of the challenges in building and developing the post-discharge model is determining the time period during which post-discharge data should be collected and used for prediction. In this work, we used a random day model. In practice, a post-discharge model can be executed daily to identify high-risk patients, no matter when they were discharged.

Our analysis demonstrates that risk changes over time, especially during the first three months post discharge. In particular, the first 7 to 10 days are the times when frail patients are at high risk of encountering a negative outcome. A decline of the prediction performance over time suggests a variation in recovery stages for healthy patients and a difference in the risk factors for patients who died or were readmitted within the same time frame. For example, the prediction of hip fracture outcome had the lowest performance between 14 and 21 days. Several factors could cause this such as various pre-existing medical problems, the degree of activity, and the ability to attend follow-ups. Further analysis could be done to identify the change in risk factors (e.g. post-discharge complications), and locate the period of time when those factors are correlated with readmission.
Although accurate prediction is crucial, healthcare providers need to identify the values of features that are associated with a higher probability of negative outcome. Locating such values may assist clinicians in understanding the reasons behind a negative outcome, as well as identify early signs of complications. Direct intervention could be applied to lessen risks through outpatient appointments or home visits. In our risk model, the features can be categorized into clinical factors and clinical resources. Both types of features influence predictions at different levels. Specifically, the amount of clinical resources allocated to treat patients before admission and after discharge could identify the patients who have a high probability of encountering a negative outcome. Similarly, the patients who utilized more treatment resources exhibit a higher probability of experiencing a negative outcome except for post-discharge follow-up utilization. A very low utilization value implies the existence of barriers preventing the patient from going to see their healthcare provider. Even low utilization values for some features, such as the number of labs ordered during admission, are associated with negative outcome occurrence. Measuring the unexpected utilization could be leveraged as an indirect method to identify patients who are at high risk.

Conclusions

This study shows that the inclusion of data in EMRs collected post-discharge can improve the performance of negative outcomes predictions, such as readmission or death. This is notable because it shows that traditional readmission prediction models, such as LACE, which focus on information available only at the point of discharge, would benefit with updates over time (provided the information is available). We illustrated that this finding holds true for both acute (i.e., hip fracture) and chronic (i.e., congestive heart failure) patient populations. It is notable that the primary driving factors of our discovery include: 1) time out of the hospital after discharge and 2) information about the quantity of resources allocated for patients that are both physical (e.g. medication, labs, appointments) and electronic (communication, documents).
While our findings are notable, there are several limitations that we wish to acknowledge, which can serve as guidelines for further investigation. First, this study focused on sampling only one day for post-discharge information. A notable extension is evaluating risk scores on multiple consecutive days to identify changes in risk score and identify patients at-risk to apply early intervention. Second, our models neglected the semantics about clinical status that might be documented in communications between patients and their healthcare providers. For instance, patients who communicate they are in severe pain or experiencing complications from opioid medications may miss their follow-ups. Thus, information intimated to care providers may indicate signs of an impending negative outcome. As such, a notable extension to our approach would be to apply natural language processing to detect and characterize such semantics for enhanced modeling. Third, our model represented various features from disparate EHR resources. However, we did not use all patient data such as vital signs (e.g., heart rate or blood pressure), medication doses, changes in laboratory test values, etc. A future direction is to include additional relevant features while accounting for the ratio between sample size and the number of features.

Acknowledgements

This research was supported in part by an IBM Smarter Planet award and NIH grant UL1 TR000445.

References

Adapting Nielsen’s Design Heuristics to Dual Processing for Clinical Decision Support

Teresa Taft1, Catherine Staes, BSN, MPH, PhD1, Stacey Slager, MS1, Charlene Weir, RN, PhD1
University of Utah1, Salt Lake City, Utah

Abstract
The study objective was to improve the applicability of Nielson’s standard design heuristics for evaluating electronic health record (EHR) alerts and linked ordering support by integrating them with Dual Process theory. Through initial heuristic evaluation and a user study of 7 physicians, usability problems were identified. Through independent mapping of specific usability criteria to support for each of the Dual Cognitive processes (S1 and S2) and deliberation, agreement was reached on mapping criteria. Finally, usability errors from the heuristic and user study were mapped to S1 and S2. Adding a dual process perspective to specific heuristic analysis increases the applicability and relevance of computerized health information design evaluations. This mapping enables designers to measure that their systems are tailored to support attention allocation. System 1 will be supported by improving pattern recognition and saliency, and system 2 through efficiency and control of information access.

Introduction
The American Medical Association (AMA) recently issued a statement calling for improved design of health information technology (HIT) systems that reduces cognitive workload, stating that, “health IT is misaligned with the cognitive and workflow requirements of medicine”. Problems with healthcare information display have been shown to reduce a user’s ability to detect changes, cause delays in finding information, increase working memory demands, lead to confusion, and lead to confidence in wrong judgments. Few electronic health records (EHRs), currently in use, improve comprehension. In drawing the conclusion that many HIT systems lack needed usability, the Institute of Medicine (IOM) called on the governing boards to require HIT, used in the diagnostic process, to demonstrate usability, incorporate human factors knowledge, and provide clinical decision support. In response to concerns and recommendations, the Office of the National Coordinator issued new meaningful use regulations, which took effect in 2016, requiring usability testing and reporting of HIT systems seeking certification.

Jakob Nielsen, a leader in IT usability design and evaluation, created a taxonomy of design heuristics that could be used by usability experts to inspect systems, which he advises should be adapted for any specific interface. In addition to this heuristic evaluation, Nielsen recommended follow-up user testing, which may include observations, ‘think aloud studies’, and interviews. Nielsen demonstrated that testing three to five users at two sites per iteration is sufficient for identification of most usability issues. Think aloud methodology, based on protocol and analysis techniques of Ericsson and Simon (1980), is also informative and is widely used in usability and interface design. Although more recent iterations of usability heuristics have been created, they all build upon Nielsen’s work.

Psychologists have been discussing two different styles of thinking for at least the last 50 years, however the terms System 1 and System 2 were first used by Stanovich and West and were more recently popularized and illustrated by Kahneman in his 2011 book, Thinking Fast and Slow. In this paper, we will use the terms S1 and S2 to discuss the two ways of thinking. S1 refers to the cognitive process operating when patterns are quickly recognized and assumptions are effortlessly made regarding the situation. Automatic, response is determined without conscious awareness. S1 consists of layers of associative memory organized into patterns or schemas that when activated produce the experience of instant pattern recognition and supports rapid automatic responses. Difficult tasks, made up of many small decisions, that are well learned to the point of expertise, such as reading, talking, and driving are easily activated in complete units by S1 processes. Recognition, “thinking” and decision-making feels effortless and requires very little cognitive attentional resources. Because S1 cognitive processing involves spreading activation through tightly associative networks, common biases occur. On the other hand, S2 cognitive processing occurs when attention is required for a task, either because it is novel, complicated, or involves a significant risk. Focused attention, as is required by S2 cognitive processes, is quite effortful and limited, rather like a flashlight beam in a very dark room. As a result, S2’s slower, more deliberative reasoning processes can easily be interrupted, overwhelmed, or exhausted. Anything that draws attention away from S2 focus can displace the information in working memory. The default cognitive processing mode is S1, but the common cognitive activities of S2, such as planning and simulating possible outcomes, as occurs during intentional decision-making, are also usually in play.
Many studies show that the two systems operate in tandem. Context affects the differential reliance on the two systems with S1 functions dominating when situations are overwhelming, emotional, or urgent and S2 dominating during highly analytical ‘hyper-focused’ thinking.

In addition to recommending usability testing, the Institute of Medicine has reported that electronic medical systems provide the opportunity to reduce diagnostic error by supporting clinician’s dual cognitive processes, System 1 and System 2, by providing information in ways that enhance pattern recognition and provide rapid access to specific data and tools that assist in cognitive processing needed for medical decision making. Designing effective decision support for clinicians in the complex working environments of healthcare requires understanding of these two modes, their relatively different information needs and the differential mechanisms required for providing support. Consideration should be given to supporting S1 pattern recognition, in signaling S2 that attention is needed, and providing S2 with timely, accurate, understandable information to support an efficient, accurate analysis. Methods of measuring the alignment of S1 and S2 with HIT systems design are lacking.

Description of a Health IT Problem
At University of Utah Healthcare (UUHC), major efforts are underway to improve the delivery of healthcare, through standardizing care processes, implementing and evaluating clinical decision support, and measuring the value of care. To improve the efficiency with diagnostic testing and management of acute knee pain in the outpatient setting, care process models were implemented, along with alerts to be delivered at the point of care and order sets to guide diagnostic and referral practices. The primary goal of the new acute knee pain care process model was to reduce the use of advanced imaging (i.e., MRIs) without first obtaining a radiograph (i.e. x-ray). Clinical decision support resources were implemented with Epic to support this goal. Specifically, two alerts could fire as a physician began ordering in the EMR. One alert informed the physician that a new knee pain order set was available as “order entry” was clicked; a second alert, which only fired if the physician attempted to order an MRI and the system could not detect a knee x-ray for the patient in the preceding 6 months, advised the clinician that x-ray should be obtained before MRI. The alerts included links to a flowchart and an order set that could be opened from the alerts. The order set consisted of a short list of ordering options with radio buttons for selection.

Usage data indicated that the alerts were canceled 95% of the time (n=1170/1236) over a 12 month span (unpublished data). Analysis of ordering practices showed that the clinical pathway resources did not achieve their objective to reduce the proportion of MRIs ordered without a prior x-ray. Before expanding or abandoning alerts, it is important to evaluate the reasons for the ineffective clinical alerts and associated smart order set, and identify areas of improvement in design that could be applied to future evaluation and design of HIT systems for encouraging compliance with best practices. We hypothesized that the lack of usage of the alerts concerned usability problems with the alerts, and dual processing considerations. To explore the S1 and S2 principles described above, we performed a usability evaluation of the active alerts and associated order set which were implemented in a limited set of outpatient clinics at UUHC.

Objectives
The purpose of this paper is to explore the relationship between common usability metrics and a dual process perspective. Mapping usability evaluations to the dual cognitive processes may be an important step in improving the application of usability testing in the design and implementation of EHRs in order to better support cognition in medical decision-making. Therefore, our specific objectives were to: a) conduct a usability evaluation of active alerts and the associated order set, and b) map the usability violations to the dual cognitive processes.

Methods
Qualitative and descriptive methods were used to perform this study. The usability evaluation included heuristic analysis, simulation with think aloud, and semi-structured interviews. The user study was conducted at two primary care outpatient clinics that are part of the University of Utah Healthcare enterprise, located in Salt Lake City Utah. The two clinics both use the Epic EHR. Family practice physicians were asked by their supervisor to participate in this quality improvement study. Both resident and experienced physicians were recruited in order to reflect actual practice (N=7).

Phase 1: Heuristic evaluation of alert-related artifacts
Three authors (TT, CS, and SS) independently performed a heuristic evaluation, interacting with the system, two knee pain alerts (BPAs), and an associated order set (SmartSet). Problems encountered were logged and categorized.
using Nielsen’s ten usability heuristics. The two ‘Error’ heuristics were combined for efficiency.

**Phase 2: User study using simulation and observation and semi-structured interview**

A think aloud simulation user study was conducted at two family practice clinics where the HIT had been deployed in a pilot implementation for the preceding year. We developed a knee pain vignette, which was evaluated and refined based on feedback from three primary care providers. Physicians who agreed to participate were presented with the patient vignette and were given instructions on how to ‘think aloud’ while they entered orders into the EHR. For example, we instructed: “I want you to speak out loud any thoughts that you have as you are using the EHR system to enter orders for the patient. For example, you might say, “I want to enter a visit diagnosis, so I click here.” Describe what you would do.” Or you might say, “What just happened there? I didn’t expect that,” or “I don’t know what this is for,” or “I really like this...” We are looking for things you like, things you don’t like, what you would change.

The study duration was 30-40 minutes per participant. Physicians accessed a test version of the EHR system on a laptop computer in a conference room at their clinic site. Each physician responded to two alerts. Response to the alerts and the reasons physicians gave for canceling them were recorded.

The patient vignette was the following: A 24 y/o male with no past medical history comes to your office after sustaining a knee injury while playing soccer 2 days ago. He reports that he collided on the field with another player. His right knee came into contact with the player’s knee. He remembers his knee being 'pushed back'. It may also have been twisted. He doesn't recall any discrete pop. He fell down and was unable to bear weight at the time. He was helped from the field and didn't return to play. Shortly after, his knee had some bruising of the kneecap, swelling, and was stiff. He borrowed a friend’s crutches, wrapped the knee with an ace wrap, and iced it for the next 24 hours. He also took Aleve, 1 tablet every 12 hours. The pain has not improved, and he is still needing the crutches. The ‘stiffening' has worsened, and he feels like it is too painful to walk on. He hasn't moved it much, but doesn't think it locks up. He didn't go to work today or yesterday. He works as a computer programmer.

The findings on the physical exam were:

- Pt walks without bearing weight on affected knee, which is flexed to 30 degrees wrapped.
- Pt has bruising of his kneecap, with diffuse tenderness, most notably over mid patella.
- There is a joint effusion.
- Passive range of motion is limited to 70 degrees of knee flexion. Knee extension and flexion strength is normal.
- Joint line is non-tender. No tenderness of tibia/fibula.
- LCL/MCL are within normal limits.
- ACL/PCL testing is not possible secondary to his pain.

The physician was given the clinical vignette with the following instructions: “Imagine a patient has come to your office after injuring his knee. I have entered a chief complaint of ‘Acute Traumatic Knee Pain’ in advance of the visit. Imagine that you have examined him and found the information in this vignette and you are now ready to write orders. Please “Think Aloud” as you enter orders for the patient in EPIC as you would if you were seeing him in clinic.”

The investigators recorded usability issues that were observed or voiced in the think aloud simulation, recording the EHR location and the description of the Problem. Immediately following the ‘think aloud’ observation, the investigators asked participants the following 12 guided questions and manually recorded the responses. (Table 1)

**Table 1. Guided questions.**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What does the purpose of this alert appear to be?</td>
<td></td>
</tr>
<tr>
<td>2. Why do you think you might cancel this alert?</td>
<td></td>
</tr>
<tr>
<td>3. How much of this alert would you read if it fired during a real clinical visit?</td>
<td></td>
</tr>
<tr>
<td>4. Is anything unclear? What?</td>
<td></td>
</tr>
<tr>
<td>5. What do you expect to happen if you accept the alert?</td>
<td></td>
</tr>
<tr>
<td>6. What do you expect to happen if you cancel the alert?</td>
<td></td>
</tr>
<tr>
<td>7. What happens to comments that are entered? Who might use those comments?</td>
<td></td>
</tr>
</tbody>
</table>
Following the user study, usability problems were classified by heuristic sub-categories developed by the authors, based on a review of the literature and the prior heuristic evaluation. An additional category of transparency was added to address physician concerns regarding accuracy of the information. The subcategory list was expanded as needed for coding of problems that didn’t fit well in the initial subcategories. (See appendix A)

**Phase 3: Dual process coding**
Problems found in the usability assessment were translated into specific heuristic subcategories. An initial coding protocol was created by the authors (TT and CW) based on their knowledge of the literature about dual processing and was subsequently developed iteratively through several rounds of independent coding by the authors. Definitions for S1 and S2 were identified by a search of the literature and refined through discussion. Once the coding protocol was established and a final level of acceptable inter-rater reliability was reached (92.9%), TT independently mapped the identified usability problems to the Dual Cognitive processes (S1 and S2). The protocol guideline for coding usability errors and subcategories is listed in Table 2.

**Table 2 Description of System 1/System 2 Coding Protocols**

<table>
<thead>
<tr>
<th>System 1: Does this feature or quality support subconscious, instantaneous knowledge/response?</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Knowledge obtained through expert language, reading, without any effortful analysis</td>
</tr>
<tr>
<td>- Automatic recognition of differences by the sensory system (auditory alarm, visual difference…)</td>
</tr>
<tr>
<td>- Instantly knowing something is correct versus incorrect</td>
</tr>
<tr>
<td>- Instantly knowing what to do or where to look</td>
</tr>
<tr>
<td>- Rapid pattern recognition, without any conscious effort</td>
</tr>
<tr>
<td>- Color and shape that consistently convey meaning</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System 2: Does this feature or quality support conscious intentional analysis and decision-making?</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Prevents or causes problems that must be solved</td>
</tr>
<tr>
<td>- Requires conscious thought or analysis to understand information or to determine the next action</td>
</tr>
<tr>
<td>- Requires time and focus to understand or decide on the next action</td>
</tr>
<tr>
<td>- Focuses attention (or does not divert attention from tasks)</td>
</tr>
</tbody>
</table>

**Results**

The evaluations identified 36 usability errors in the alerts and associated order set. Half of the errors (n=19, 52.7%) were discovered through heuristic analysis. Other errors were suspected, but confirmation with users was needed. The remaining errors (n=17, 47.2%) were confirmed or discovered during the user study. Two-thirds of the identified usability errors (n=23, 64%) were mapped to interference with S2 cognitive processes and one-third of the errors (n=13, 36%) were mapped to interference with system 1 processes.

Table 3 includes the list of errors classified by heuristic, and mapped to the relevant cognitive process. During the user study, five types of response to the alerts were observed (Table 4). Physicians attempting to accept the alerts consistently had difficulty managing the alert functions. We documented ten reasons that physicians gave for canceling alerts (Table 5). Physicians gave more than one reason. Seven reasons that alerts were canceled were associated with efficiency and three reasons were associated with trust in the accuracy or function of the alert.
Table 3. Description of usability errors mapped to the Dual Cognitive Processes.

<table>
<thead>
<tr>
<th>Nielsen’s Usability Heuristic</th>
<th>Source</th>
<th>Quote or Description of Problem</th>
<th>Cognitive system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aesthetic and minimalist design</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>x</td>
<td>Additional information on the same line as the alert advisory diverts attention from the focus</td>
<td>S1</td>
</tr>
<tr>
<td>A11</td>
<td>x</td>
<td>The requested information in the alert was for usability analysis. Collecting usability data from within an EHR alert is not appropriate.</td>
<td>S2</td>
</tr>
<tr>
<td>Consistency &amp; Standards</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>x</td>
<td>Names on hyperlinks do not match the linked document title.</td>
<td>S1</td>
</tr>
<tr>
<td>C6</td>
<td>x</td>
<td>The link to the pathway guidance becomes physically separated from associated information when more than one SmartSet opens.</td>
<td>S1</td>
</tr>
<tr>
<td>C9</td>
<td>x</td>
<td>“It would be clearer if the MRI results were in boxes like the x-ray results are.”</td>
<td>S1</td>
</tr>
<tr>
<td>Error Prevention, diagnosis &amp; recovery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>x</td>
<td>When ordering imaging from the smart order set, the preset date is outside the limits allowed by the system. This error requires 4 clicks to resolve.</td>
<td>S2</td>
</tr>
<tr>
<td>E1</td>
<td>x</td>
<td>When ordering imaging, options of future, normal and standing can be selected, however only an order status of future is allowed by the system. This error requires 7 clicks to correct.</td>
<td>S2</td>
</tr>
<tr>
<td>E1</td>
<td>x</td>
<td>Typing a reason in the acknowledge reason box is an error. Free text typing is allowed in the acknowledge reason box, but results in an error that is difficult to resolve unless canceled.</td>
<td>S2</td>
</tr>
<tr>
<td>E1</td>
<td>x</td>
<td>Ordering imaging within the SmartSet produces a box with an error code that must be closed to proceed.</td>
<td>S2</td>
</tr>
<tr>
<td>E2</td>
<td>x</td>
<td>“Patient’s often receive care outside of the University system, so BPAs specific to the patient are often not accurate.”</td>
<td>S2</td>
</tr>
<tr>
<td>E3</td>
<td>x</td>
<td>Nothing happens when the magnifier is clicked. No help or information is given.</td>
<td>S1</td>
</tr>
<tr>
<td>E3</td>
<td>x</td>
<td>Alternative methods for ordering, which bi-pass the alerting system are available and are frequently used</td>
<td>S1</td>
</tr>
<tr>
<td>Flexibility &amp; Efficiency of Use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>x</td>
<td>“The imaging orders should have an expected date of ‘today’ pre-selected so that I don’t always have to enter a date.”</td>
<td>S2</td>
</tr>
<tr>
<td>F4</td>
<td>x</td>
<td>“Every time I accept a BPA, cumbersome “order sets” are pulled up that are laborious and frustrating, so I scan the BPAs, then I always cancel them.” (Physician demonstrated how many boxes must be unchecked)</td>
<td>S2</td>
</tr>
<tr>
<td>F4</td>
<td>x</td>
<td>“When I use a SmartSet, I always have to associate the orders with the visit diagnosis twice. This should be fixed so that I only have to complete this step once.”</td>
<td>S2</td>
</tr>
<tr>
<td>F4</td>
<td>x</td>
<td>“SmartSets are a hassle when items are preselected. I wouldn’t use it because it requires too much clicking.”</td>
<td>S2</td>
</tr>
<tr>
<td>Help and Documentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>x</td>
<td>“I tried to tell it I would do what it was asking. It didn’t help me.”</td>
<td>S2</td>
</tr>
<tr>
<td>H1</td>
<td>x</td>
<td>When using the SmartSet to place an order, the physician couldn’t figure out how to add the diagnosis.</td>
<td>S2</td>
</tr>
</tbody>
</table>
Table 3. Description of usability errors identified and mappings to the Dual Cognitive Processes (continued).

<table>
<thead>
<tr>
<th>Nielsen’s Usability Heuristic</th>
<th>Source</th>
<th>Quote or Description of Problem</th>
<th>Cognitive system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match between System and the Real World</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1 x x</td>
<td>“The links in the alert are not clearly labeled. I am not sure what they are.”</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>M1 x</td>
<td>Physicians did not know what the label “0 out of 4” meant</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>M1 x</td>
<td>“I have no idea what ‘follow-up action taken’ means”</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>M4 x</td>
<td>Orders the physician normally writes are not available in the smart order set.</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>M5 x</td>
<td>“This is for two separate processes: (1) diagnosis, and (2) referral. I would prefer them to be separate.”</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>M6 x</td>
<td>“Very obvious things are included leading to unnecessary clutter”</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>M6 x</td>
<td>“I would like the SmartSet to have fewer options”</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>Recognition rather than recall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2 x</td>
<td>Medical decision criterion was confusing to physicians.</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>R5 x</td>
<td>The guidelines &amp; smart order set cannot be viewed at the same time.</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>Transparency*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T7 x x</td>
<td>“I don’t know why the pathway recommends 4-x-rays instead of 3.”</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>User Control &amp; Freedom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1 x</td>
<td>“SmartSets are difficult to use or turn off.”</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>U3 x</td>
<td>“I get skeptical of links in the EMR because I worry it will interrupt what I am doing.”</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>U3 x x</td>
<td>“SmartSets are a hassle when multiple SmartSets open together. I don’t want to deal with all those issues at the same time.”</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>U6 x x</td>
<td>Multiple alerts appear together with only one cancel and one accept button. It appears that they all must be accepted or all canceled.</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>U11 x x</td>
<td>Pop-up alert interrupts the ordering process to ask user to order from the smart order set</td>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>Visibility of System Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V4 x x</td>
<td>Physician didn’t know to click on headings to open nested hidden order options within the SmartSet.</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>V6 x</td>
<td>“If it’s a SmartSet, then I probably don’t want to use it. It may put in a weird diagnosis, not one I usually use.”</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>V6 x</td>
<td>“When I see an unfamiliar BPA, I cancel it, because I don’t know what it will do.”</td>
<td>S1</td>
<td></td>
</tr>
</tbody>
</table>

*Transparency Heuristic was added by the authors to address physician comments regarding accuracy of information.

Table 4. Types of responses to the alerts associated with acute knee pain observed during the user study

<table>
<thead>
<tr>
<th>Type of response</th>
<th>Description of actions observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignore</td>
<td>Automatically cancels the alert without ‘seeing it’</td>
</tr>
<tr>
<td>Skim</td>
<td>Reads only 3-4 words in the title and cancels</td>
</tr>
<tr>
<td>Read</td>
<td>Reads and cancels</td>
</tr>
<tr>
<td>Attempt</td>
<td>Reads, attempts response, and cancels</td>
</tr>
<tr>
<td>Persist</td>
<td>Reads, responds-correcting mistakes, and accepts</td>
</tr>
</tbody>
</table>
Table 5. Reasons that physicians cancelled alerts

<table>
<thead>
<tr>
<th>Reason provided by physician</th>
<th>Efficiency</th>
<th>Trust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too much clicking/time involved in use of the alert</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Uncertainty about functions of the alert</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Uncertainty about who is the audience for the “comments”</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Alert functions disrupted workflow and thought process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distrust of patient specific information provided in alerts</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Accepting alerts creates long lists of order options</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple alerts open together, but there is only one cancel and one accept button</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Habitual canceling of all alerts (“I always cancel alerts”)</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Alert blindness (“I didn’t see it” referring to the alert that they had just cancelled)</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>The alert functions were not viewed as helpful.</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

Discussion

This study integrated two approaches to assessing usability and design of HIT. The first approach was a traditional usability study. The second was a mapping of the usability issues in the context of its support or interference with dual cognitive processes. Investigators identified five ways that physicians responded to ineffective alerts, reflecting various levels of attention, effort, and deliberation. Ten reasons for cancelling of alerts were also tracked with seven related to efficiency, and the remaining three reasons related to trust in the information or function of the alert. Alerts are intended to focus S2 cognition on something important, however, the reasons given for cancelling the alerts provided evidence that the alerts were not supporting the needs of S2 cognitive processes. These needs included the ability to focus without distraction, to evaluate the validity of information, and to efficiently accomplish goals. In mapping the usability errors of a clinical pathway within an EHR to a S1/S2 model, we found that nearly 2/3 were problems with supporting S2 logical decision processes. Adaptations of appendix A may prove to be useful for usability evaluation and specific HIT design recommendations to support S1 and S2 processes.

Distrust in the accuracy of recommendations provided by the system and the skill with which physicians avert the interruptions to their cognitive processes through cancellation of alerts stand out in this study. Although the alert firing algorithm appeared on the surface to be timely and targeted, physicians spoke with the clicker that these alerts were not useful, as they almost always cancelled them. Physicians adapted to a complex environment by focusing on proven ‘targets’ and away from less productive ‘noise’. To avoid training physicians to view alerts as ‘noise’, consideration should be given to implementing accuracy and usability testing prior to any alert deployment and post deployment monitoring. Physicians stated that previous experience with alerts that provided erroneous information reduced their trust and reliance on alerts. Consideration should be given to disabling of alerts that do not meet an accuracy threshold with requirements determined by likelihood & severity. Methods that increase transparency, including metadata and links to source information, may reduce uncertainty and increase trust in alerts. In developing & evaluating CDSS, including an additional heuristic of ‘transparency’ with Nielsen’s design heuristics may promote design that increases physician trust in information.

Physicians reported difficulty in managing the functionality of EHR subsystems that they rarely used. Evaluation of the alerts showed that they were more complex than was necessary and poorly matched with the users mental models and workflow. This poor match contributed to difficulty physicians experienced in predicting the results of interaction with the system. Simplification of alerts and improved labeling may help to reduce high cancelation rates. A workflow analysis prior to design could have revealed that ordering decisions are made prior to clicking the order tab. Support for cognitive processes at the point of decision-making may be far superior to corrective alerting, which comes sometime after a decision has been made.

The focus of information technology design has been primarily on functionality (the ability of the system to meet functional requirements, once the user is trained to use the system), however the failure of many functional systems demonstrates the need for usability to be addressed in the earliest stages of clinical decision support system (CDSS) design, rather than as an afterthought. Prototype driven CDSS design shows great promise in allowing usability evaluation to take its place at the forefront of the CDSS train, rather than its usual position as the often forgotten caboose. Careful evaluation before deployment of alerts is paramount if we are to avoid training physicians to cancel all alerts. Design methods should include iterative usability testing conducted by usability experts who can see the project with fresh eyes because they did not build it, whenever possible.
Limitations
This study investigated methods for mapping the dual cognitive processes to usability assessment. Problems were identified in an alert that was almost always canceled, creating an opportunity to discover many usability errors. Specific errors identified in other systems will necessarily be different and the percentage of problems that affect S1 and S2 can be expected to vary. Although usability errors were mapped to not supporting one system or another, resolution of a problem created by the usability error was not addressed and at times, the alternate cognitive process may manage the task, or both cognitive systems may be involved in this process. The evaluation was only applied to one setting. Validation and generalizability of the methods will be determined by future work. In addition, these methods should be adapted for specific use.

Conclusion
Support for the intuitive and analytical dual cognitive processes is needed in CDSS design. ‘Alert Blindness’ attributable to cognitive focusing on reliable ‘targets’ rather than ‘noise’ must be addressed if alerts are to be effective in supporting deliberative S2 clinical decision-making. Standard alert monitoring, usability testing, and reporting are needed for alerts embedded in the electronic ordering systems. Turning off alerts identified as ineffective or inappropriate could restore their usefulness as a ‘fire-alarm’ redirect of S2 attention to vital concerns. Improved accuracy standards, and transparency of system delivered information along with greater simplicity and consistency in technology functions are needed to enable S1 to automatically manage these activities, freeing the working memory of S2 for needed analytical processes.

Improved design of systems to present integrative views that support S1 automatic processing and to support clinician control of their own S2 attention allocation is needed. Standard measures to assess and report CDSS effect on cognition should be adopted to support design, evaluation and comparison of clinical pathways within a healthcare system. The mapping of usability issues to support for the dual cognitive systems, as was done in this study, is one approach that may prove useful for designing HIT systems that more effectively support the diagnostic process.

References

Appendix A. Usability sub categories used in evaluation of clinical pathway support of cognitive systems

<table>
<thead>
<tr>
<th>Code</th>
<th>Modified Nielsen’s Heuristics</th>
<th>Cognitive System</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Aesthetic and minimalist design</td>
<td>A1 Sequences of action of clear</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A2 Presentation of information is efficient and concise</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>A3 Font sizes are large enough for easy reading</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A4 Text has high contrast from background</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A5 Picture allows rapid confirmation that the correct patient or pathway is being viewed</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A6 Titles / naming is concise and unique (condition is listed before modifiers)</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>A7 Need for scrolling is obvious if necessary</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A9 No problems for colorblind</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A10 Visual display of information, such as charts and graphs aid interpretation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A11 Only essential information is collected</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>A12 Use of color aids comprehension</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A13 User can quickly see how to use the clinical pathway document</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>A14 First word in menu choices is most important</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A15 Flashing text or icons are avoided</td>
<td>1/2</td>
</tr>
<tr>
<td>C. Consistency and standards</td>
<td>C1 Consistent use of color &amp; shape to communicate context</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C2 Layout and position has spatial consistency</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C3 Font use is consistent (levels)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C4 Language conventions, terminology, spelling, capitalization</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>C5 Keywords, labels are consistent across screens &amp; pathway tools</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C6 Standard position &amp; usage of elements (e.g., blue underlined text for hyperlinks)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C7 Diagram and order-sets map</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>C8 Degree of detail is consistent</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>C9 Shapes have consistent meaning and follow conventions</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C10 Diagrams progress from left to right and top to bottom when possible</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C11 Flowchart lines do not cross</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C12 Arrows indicate the direction of the flow for all lines in the diagram</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C13 Actions in flow charts are written verb first</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>C14 Menu choices are titles in lower level menus</td>
<td>1</td>
</tr>
<tr>
<td>E. Error prevention, recognition, diagnosis &amp; recovery</td>
<td>E1 Interface makes errors impossible</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>E2 Information presented is reliable</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>E3 The system functions as intended</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>E4 Error messages use clear language and are precise (not system error #8739)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>E5 Error messages are polite (not “fatal”, “failure”, “illegal”...)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>E6 Error messages indicate the recovery process</td>
<td>2</td>
</tr>
<tr>
<td>F. Flexibility and efficiency of use</td>
<td>F1 Allows users to tailor frequent actions</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>F2 Shortcuts for experienced users</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>F3 Shortcuts and tailoring are assisted by the system—simple to implement</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>F4 Tedium sequences are avoided</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>F5 Typing/ clicking requirements are minimized</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>F6 Use of the system is intuitive and does not require training to use it</td>
<td>1</td>
</tr>
</tbody>
</table>

H. Help and documentation
<table>
<thead>
<tr>
<th></th>
<th>Instructions are visible or easily accessed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>H2</td>
<td>Clearly focuses on the user’s task</td>
<td>2</td>
</tr>
<tr>
<td>H3</td>
<td>Lists concrete steps to be carried out</td>
<td>2</td>
</tr>
<tr>
<td>H4</td>
<td>Is concise &amp; easy to follow</td>
<td>2</td>
</tr>
<tr>
<td>H5</td>
<td>Clear prompts for data entry</td>
<td>1</td>
</tr>
<tr>
<td>M. Match between system and the real world</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>The system speaks the user’s language</td>
<td>1</td>
</tr>
<tr>
<td>M2</td>
<td>Information appears in a natural and logical order</td>
<td>2</td>
</tr>
<tr>
<td>M3</td>
<td>Process matches workflow</td>
<td>1</td>
</tr>
<tr>
<td>M4</td>
<td>Available options match needs of the user</td>
<td>2</td>
</tr>
<tr>
<td>M5</td>
<td>Information is delivered at the right time to the right person</td>
<td>2</td>
</tr>
<tr>
<td>M6</td>
<td>Degree of detail is appropriate for the intended user</td>
<td>2</td>
</tr>
<tr>
<td>R. Recognition rather than recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>Annotations and footnotes detail information about decisions or processes</td>
<td>2</td>
</tr>
<tr>
<td>R2</td>
<td>Decision criterion is clearly stated</td>
<td>2</td>
</tr>
<tr>
<td>R3</td>
<td>Links to supporting documents (evidence base, policies...) are active</td>
<td>2</td>
</tr>
<tr>
<td>R4</td>
<td>Headings conveys essential information in the first 4 words</td>
<td>1</td>
</tr>
<tr>
<td>R5</td>
<td>All necessary information is visible</td>
<td>2</td>
</tr>
<tr>
<td>R6</td>
<td>Exit points show when the patient should leave the care pathway</td>
<td>2</td>
</tr>
<tr>
<td>T. Transparency*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>Developers/stewards are listed</td>
<td>2</td>
</tr>
<tr>
<td>T2</td>
<td>Intended user is clearly identified</td>
<td>1</td>
</tr>
<tr>
<td>T3</td>
<td>Goals of the pathway are clearly stated</td>
<td>2</td>
</tr>
<tr>
<td>T4</td>
<td>Date of pathway development or last update is clear</td>
<td>1/2</td>
</tr>
<tr>
<td>T5</td>
<td>Clinical rationale is associated with specific best practices evidence</td>
<td>2</td>
</tr>
<tr>
<td>T6</td>
<td>Contact information is listed with an invitation to give feedback for improvement</td>
<td>2</td>
</tr>
<tr>
<td>T7</td>
<td>References are cited to enable assessment of relevance and quality of guidelines</td>
<td>2</td>
</tr>
<tr>
<td>T8</td>
<td>Source data is available for verification of information</td>
<td>2</td>
</tr>
<tr>
<td>T9</td>
<td>Accuracy rates of system provided patient data is reported</td>
<td>2</td>
</tr>
<tr>
<td>U. User control and freedom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>Users can exit quickly.</td>
<td>1</td>
</tr>
<tr>
<td>U2</td>
<td>Undo and redo are supported</td>
<td>2</td>
</tr>
<tr>
<td>U3</td>
<td>Users are initiators of action, not responders</td>
<td>2</td>
</tr>
<tr>
<td>U4</td>
<td>Users can quickly insert themselves at the point they would like in the pathway</td>
<td>2</td>
</tr>
<tr>
<td>U5</td>
<td>Alerts can be turned off</td>
<td>2</td>
</tr>
<tr>
<td>U6</td>
<td>Users are not forced by the system to choose all or none</td>
<td>2</td>
</tr>
<tr>
<td>U7</td>
<td>Options are shown. What can be done next is clear</td>
<td>2</td>
</tr>
<tr>
<td>U8</td>
<td>References have active links to support access</td>
<td>1</td>
</tr>
<tr>
<td>U9</td>
<td>Available options match needs of the user.</td>
<td>2</td>
</tr>
<tr>
<td>U10</td>
<td>The care pathway is easily accessed by intended users when needed or desired</td>
<td>2</td>
</tr>
<tr>
<td>U11</td>
<td>Alert importance justifies the level of interruption</td>
<td>2</td>
</tr>
<tr>
<td>U12</td>
<td>Users can enter data in any order they desire</td>
<td>2</td>
</tr>
<tr>
<td>U13</td>
<td>Users can easily move forward or backward in successive screens</td>
<td>1</td>
</tr>
<tr>
<td>V. Visibility of system status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1</td>
<td>Options are shown. What can be done next, where the user can go</td>
<td>1</td>
</tr>
<tr>
<td>V2</td>
<td>Process flow is clear. Decisions, and next steps are easy to follow</td>
<td>2</td>
</tr>
<tr>
<td>V3</td>
<td>Tasks have a clear beginning and end</td>
<td>1</td>
</tr>
<tr>
<td>V4</td>
<td>Collapsible lists are clearly indicated when closed</td>
<td>1</td>
</tr>
<tr>
<td>V5</td>
<td>Links are easily identified</td>
<td>1</td>
</tr>
<tr>
<td>V6</td>
<td>Results of interactions with the system are predictable</td>
<td>1</td>
</tr>
<tr>
<td>V7</td>
<td>Feedback clearly communicates changes implemented by an action</td>
<td>1</td>
</tr>
<tr>
<td>V8</td>
<td>Labels &amp; headings clearly convey contents for data sets</td>
<td>1</td>
</tr>
<tr>
<td>V9</td>
<td>Visual feedback shows where the cursor is</td>
<td>1</td>
</tr>
<tr>
<td>V10</td>
<td>Visual indication of what has been selected (i.e. bold, color, check...)</td>
<td>1</td>
</tr>
</tbody>
</table>

*Transparency Heuristic was added by the authors to address physician concerns regarding accuracy of information.
Automating Guidelines for Clinical Decision Support: Knowledge Engineering and Implementation

Geoffrey J. Tso, MD1,2, Samson W. Tu MS2, Connie Oshiro PhD1, Susana Martins MD MSc1, Michael Ashcraft MD1, Kaeli W. Yuen1, Dan Wang PhD1, Amy Robinson PharmD1, Paul A. Heidenreich MD MS1,2, Mary K. Goldstein MD MS1, 2

1VA Palo Alto Health Care System, Palo Alto, CA; 2Stanford University, Stanford, CA

Abstract

As utilization of clinical decision support (CDS) increases, it is important to continue the development and refinement of methods to accurately translate the intention of clinical practice guidelines (CPG) into a computable form. In this study, we validate and extend the 13 steps that Shiffman et al.5 identified for translating CPG knowledge for use in CDS. During an implementation project of ATHENA-CDS, we encoded complex CPG recommendations for five common chronic conditions for integration into an existing clinical dashboard. Major decisions made during the implementation process were recorded and categorized according to the 13 steps. During the implementation period, we categorized 119 decisions and identified 8 new categories required to complete the project. We provide details on an updated model that outlines all of the steps used to translate CPG knowledge into a CDS integrated with existing health information technology.

Introduction

Clinical decision support (CDS) has been shown to improve the quality of healthcare delivered.1,2 With rising utilization of an electronic healthcare record (EHR) in clinical practice, there are increasingly more opportunities to incorporate CDS into additional aspects of the healthcare workflow. In order to improve the effectiveness, adoption, and sustainability of CDS implementations, there has been an increased interest in implementing systems that encapsulate more medical knowledge.3,4

Clinical practice guidelines (CPG) continue to hold promise as a source of evidence-based medical knowledge, and there has been significant work aiding the translation of this knowledge into software systems.5-7 However, guidelines often contain gaps and ambiguities, making it difficult to implement or operationalize the knowledge into a computable format.5 These shortcomings result in varied interpretations of the same recommendation and decreased adoption.8,9 With the aid of clinical experts, it is possible to interpret and complete the knowledge required to implement intentions of the guidelines completely; however, this process is not yet standardized and there are many aspects to consider for a successful end product.5

Shiffman et al. formalized steps for translating the knowledge contained in guideline text into a computable format that could be operationalized into a CDS specification.5 Since the publication of the seminal paper, most CDS implementations must consider integration directly with another system, and knowledge modeling techniques have continued to develop. In this study, we validate and update the steps in the Shiffman model, and we extend them to apply to the full process of implementing a CDS system based on CPG knowledge. We show that, in addition to deriving CDS knowledge completely from guideline documents, implementing a CDS system often requires changes to the design specification.

Methods

Clinical Decision Support System

We conducted this study during a modified implementation of ATHENA-CDS, an automated CPG-based CDS system that was developed at the Veterans Health Administration (VHA) of the Department of Veteran’s Affairs (VA) Palo Alto Health Care System. ATHENA-CDS uses CPG knowledge encoded in Protégé, a knowledge acquisition program developed at the Stanford Center for Biomedical Informatics Research (BMIR).10 The CPG Knowledge Base (KB) and patient EHR data are processed by the BMIR EON Guideline Interpreter system, which generates patient-specific conclusions, including recommendations for management, evidence base for the recommendations, and a summary of patient data that is relevant to decision-making.11 Clinical modules were developed, modified, or updated to provide CPG recommendations, in the context of a clinical dashboard, for the
management of five chronic diseases domains: hypertension (HTN), chronic kidney disease (CKD), heart failure (HF), hyperlipidemia (HLD), and glycemic control in type 2 diabetic patients (DM).

Implementation Context

Within Veterans Integrated Service Network (VISN) 21 of the VHA, the Pharmacy Benefits Management (PBM) group has developed a primary care clinical dashboard built on top of the VA clinical data warehouse. The dashboard provides tools for clinic team members to monitor VA clinical performance measures. ATHENA-CDS was modified to generate and display guideline-based management recommendations within the clinical dashboard.

Implementation Decision Categorization

During the development of the KBs from CPGs and the integration of ATHENA-CDS into the clinical dashboard, we recorded major implementation decisions that required interdisciplinary discussion or impacted parallel or downstream project activities. Three project members (CO, SM, KWY) extracted the decisions from meeting notes and created a spreadsheet that detailed the clinical domain, date, topic, references, issue/questions, and disposition of each decision. One project member (SWT), an experienced knowledge engineer, evaluated each decision to categorize the decisions into one of the 13 steps defined in the Shiffman model (Table 1). A senior clinician (MKG) evaluated the initial categorization, and working with the knowledge engineer, developed a consensus categorization of the decisions and proposed new categories for the list of decisions that could not be categorized. The project team, which included clinicians, knowledge engineers, and system implementers, agreed upon the final list of new categories.

Table 1. Definitions of Shiffman's Steps

<table>
<thead>
<tr>
<th>Decision Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select Guidelines</td>
<td>Choice of specific guidelines and choice of specific recommendations within the selected guidelines to be implemented</td>
</tr>
<tr>
<td>Markup</td>
<td>Identification and tagging of guideline knowledge components relevant to operationalization</td>
</tr>
<tr>
<td>Atomize</td>
<td>The process of extracting and refining single concepts from the narrative text recommendations</td>
</tr>
<tr>
<td>Deabstract</td>
<td>The process of adjusting the level of generality at which a decision variable or action is described to permit operationalization</td>
</tr>
<tr>
<td>Disambiguate</td>
<td>The process of establishing a single semantic interpretation for a recommendation statement</td>
</tr>
<tr>
<td>Build Executable</td>
<td>Arrangement of the atomized, de-abstracted, and disambiguated decision variables and actions into logical statements that can be translated readily into computable statements</td>
</tr>
<tr>
<td>Verify Completeness</td>
<td>The process to make sure that each recommendation provides guidance in all situations that a clinician is likely to face</td>
</tr>
<tr>
<td>Add Explanation</td>
<td>A facility to describe the reasoning behind recommendations</td>
</tr>
<tr>
<td>Identify Origin</td>
<td>Identifying a source or origin in the clinical environment for each decision variable</td>
</tr>
<tr>
<td>Insert Recommendations</td>
<td>Identifying an insertion point in the care process for each recommended action</td>
</tr>
<tr>
<td>Define Action Type</td>
<td>Categorizing guideline-recommended activities according to predefined action types</td>
</tr>
<tr>
<td>Define Associated Beneficial Services</td>
<td>Linking action types to associated beneficial services that offer design patterns for facilitating clinical care</td>
</tr>
<tr>
<td>Design User Interface</td>
<td>Selecting and grouping user interface elements to best deliver CDS output</td>
</tr>
</tbody>
</table>

Results

We identified 119 decisions made at project meetings over a 29-month period (Examples in Table 2). During the study period, CPGs for HTN\textsuperscript{12}, DM\textsuperscript{13}, CKD\textsuperscript{14}, HLD\textsuperscript{15}, and HF\textsuperscript{16} were selected to be encoded for the KBs. Eighty of the decisions were categorized within the Shiffman model. Table 3 shows the incidence of ATHENA-CDS developmental decisions that had been categorized for each step in the model. The remaining 39 decisions were analyzed and, through consensus, placed in new categories (Table 4).
Table 2. Sample of implementation questions and issues and resulting decision

<table>
<thead>
<tr>
<th>Clinical Domain</th>
<th>Topic</th>
<th>Issue/questions</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Drug hierarchy</td>
<td>Drugs added in ad hoc fashion</td>
<td>Reorganize drugs into same hierarchy as VA NDF-RT.</td>
</tr>
<tr>
<td>Diabetes, Hyperlipidemia</td>
<td>Pregnancy exclusion</td>
<td>Difficult to identify currently pregnant women</td>
<td>Issue primary message to women ages 18-50 that pregnant women and nursing mothers are out of scope.</td>
</tr>
<tr>
<td>All</td>
<td>Medication possession ratio (MPR)</td>
<td>Include in knowledge base logic or use in post-processing in dashboard logic?</td>
<td>We are not going to use MPR to suppress recommendations, but use it to generate additional recommendations.</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>Clinical dashboard vs. Stone 2014</td>
<td>Stone: Treat patients w/ LDL&gt;190 w/o risk factors Dashboard: Out of scope</td>
<td>Align treatment goals with Pharmacy Clinical Dashboard</td>
</tr>
</tbody>
</table>

Table 3. Incidence of decisions for Shiffman’s steps

<table>
<thead>
<tr>
<th>Decision Category</th>
<th>Incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select Guidelines</td>
<td>14</td>
</tr>
<tr>
<td>Markup</td>
<td>0</td>
</tr>
<tr>
<td>Atomize</td>
<td>1</td>
</tr>
<tr>
<td>Deabstract</td>
<td>4</td>
</tr>
<tr>
<td>Disambiguate</td>
<td>0</td>
</tr>
<tr>
<td>Build Executable Statements</td>
<td>0</td>
</tr>
<tr>
<td>Verify completeness</td>
<td>19</td>
</tr>
<tr>
<td>Add Explanation</td>
<td>2</td>
</tr>
<tr>
<td>Identify Data Origins</td>
<td>2</td>
</tr>
<tr>
<td>Insert Recommendations</td>
<td>3</td>
</tr>
<tr>
<td>Define Action Type</td>
<td>5</td>
</tr>
<tr>
<td>Define Associated Beneficial Services</td>
<td>0</td>
</tr>
<tr>
<td>Specify User Interface</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 4. Incidence of new implementation categories

<table>
<thead>
<tr>
<th>New Decision Category</th>
<th>Incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconcile Multiple Guidelines</td>
<td>7</td>
</tr>
<tr>
<td>Align with Existing CDS</td>
<td>12</td>
</tr>
<tr>
<td>Adapt to Local System</td>
<td>2</td>
</tr>
<tr>
<td>Add Enhancement</td>
<td>2</td>
</tr>
<tr>
<td>Specify and Encode Knowledge</td>
<td>4</td>
</tr>
<tr>
<td>Map Terminology</td>
<td>2</td>
</tr>
<tr>
<td>Test CDS</td>
<td>8</td>
</tr>
<tr>
<td>Manage Project</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
</tr>
</tbody>
</table>

Implementation Model

From our study, eight new categories of steps in the implementation process were identified and defined (Table 5).

Table 5. Definitions of new implementation steps

<table>
<thead>
<tr>
<th>#</th>
<th>New Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reconcile Multiple Guidelines</td>
<td>Reconciliation of recommendations from different guideline sources (e.g., VA versus professional society guidelines) and guidelines for related clinical domains.</td>
</tr>
<tr>
<td>2</td>
<td>Align with Existing CDS</td>
<td>Selection and alignment of recommendations details (e.g., HbA1C targets and drugs to recommend) based on the need to be consistent with existing system.</td>
</tr>
<tr>
<td>3</td>
<td>Adapt to Local System</td>
<td>Modification of recommendations based on capabilities of local system (e.g., non-availability of data).</td>
</tr>
<tr>
<td>4</td>
<td>Add Enhancement</td>
<td>Addition of beneficial services based on the availability of local resources.</td>
</tr>
<tr>
<td>5</td>
<td>Specify and Encode Knowledge</td>
<td>Decisions about design, organization, and conventions used in representation of guideline knowledge. Encoding of knowledge base.</td>
</tr>
<tr>
<td>6</td>
<td>Map Terminology</td>
<td>Mapping terminology used in data sources to terminology used in guidelines.</td>
</tr>
<tr>
<td>7</td>
<td>Test CDS</td>
<td>Simulation of patient data to verify correctness of guideline encoding and the completeness and clinical appropriateness of system-generated recommendations.</td>
</tr>
<tr>
<td>8</td>
<td>Manage Project</td>
<td>Decisions about the scope of the project and the organization and timing of tasks.</td>
</tr>
</tbody>
</table>

1191
Most of the new steps are generalizable to all CDS implementation projects. One of the categories, Reconciliation of Multiple Guidelines, is especially derived for CDS that utilizes multiple knowledge sources. Using these observations from our study, we extend and update the Shiffman model, incorporating the eight additional steps (Table 4) that were required to implement ATHENA-CDS. Figure 1 contains the activity flow for the proposed model with 21 steps. The diagram shows that not all of the steps need to be performed in a specific order.

Figure 1. Activity diagram of updated CDS implementation model.

Here we will describe all of the 21 steps in more detail. Starred steps are new additions to the model.

1. Knowledge Preparation

1.1. Guideline Selection

Many guidelines from different organizations/writers can exist on the same topic. Choosing which guideline to implement is an important step in the development process that can affect many of the subsequent steps. In choosing which guidelines to implement, we considered several factors: validity, level of evidence, applicability in the clinical context, institutional recommendations, local practice variation, and ease of operationalization. Furthermore, within each guideline, individual recommendations also need to be selected for implementation. For some of the clinical domains, we had the choice of using recommendations or sections of recommendations from either VA/Department of Defense CPGs or those from professional societies. Whether a recommendation could be operationalized and applied to the clinical context was an important factor in the selection process.

1.2 Guideline Reconciliation *

Reconciling multiple guidelines is a new proposed step in CDS implementation. If multiple CPGs or other knowledge sources are used within the knowledge base of the CDS, differences in recommendations from each guideline must be reconciled to prevent interactions or conflicts in the CDS logic. An important feature of ATHENA-CDS is taking account of comorbid conditions when making recommendations for the index condition. For the five clinical domains in ATHENA-CDS, many of the CPGs discussed overlapping topics (eg. blood pressure
managers. Reconciliation was necessary to prevent conflicting recommendations from the CDS. We found that reconciliation between the guidelines was most important when there was a possibility of drug-drug interactions or different usages of the same medication class. For ATHENA-CDS, domain experts worked with knowledge engineers to ensure that the resulting knowledge corpus was complete and accurate.

1.3 Markup

In this Shiffman step, a guideline is converted from text into a semi-structured representation in order to identify and tag components relevant to operationalization of a CPG. In his paper, Shiffman describes an XML-modeling approach that decomposes a CPG into Guidelines Elements Model (GEM) component elements. For our purposes, an XML approach that marks up documents individually did not facilitate translation of the guideline knowledge into a Protégé knowledge base, so we took a different approach. We used a collaborative team approach for organizing the information in terms of the guideline model used in ATHENA-CDS KBs. The resulting representation facilitates the operationalization process as an intermediate view that focuses on elements important to the final knowledge encoding, such as: eligibility criteria, clinical algorithms, goals, definitions, concept hierarchies, clinical scenarios and decisions, interventions/drugs, and schemas for rating evidence quality and recommendation strength.

2. CDS Knowledge Specification and Encoding

2.1 Atomization

In order to execute the recommendations in a CPG, the recommendations in the narrative text need to be extracted and reduced to computable forms. Atomization achieves this by removing unnecessary text and identifying and normalizing the key concepts. Extraction of the guideline concepts can be performed manually or through natural language processing and text-mining tools. These concepts can then be normalized to terminologies in a custom or local library or to standardized terminologies such as Logical Observation Identifiers Names and Codes (LOINC). As an example, in the HF CPG, the selection criteria for the recommendation of an aldosterone receptor antagonist states that a patient should have:

Serum creatinine <2.5 mg/dL (or an estimated glomerular filtration rate >30 mL/min/1.73 m²) without recent worsening and serum potassium <5.0 mEq/L without a history of severe hyperkalemia

The atomization process extracted the following concepts from the statement: serum creatinine, estimated glomerular filtration rate, recent worsening, serum potassium, history of, and severe hyperkalemia. The criteria of <2.5 mg/dL, >30mLmin/m², and <5.0mEq/L were also extracted.

2.2 Deabstraction

Deabstraction is the process of converting generalized or high level concepts into specific concepts that can be operationalized. In the aldosterone receptor antagonist example, “severe hyperkalemia” needs to be converted to a more specific concept such as “serum potassium level above 1.5meq/L above normal.” While this example is relatively straightforward, other concepts can be more complex. For example, the atomized concept “metoprolol” can be further specified to individual formulations such as metoprolol tartrate or metoprolol succinate. The latter is reported to be more effective than the former in clinical trials and would be preferred. Therefore, careful attention is necessary to ensure adherence to evidence-based practice in the deabstraction process.

2.3 Disambiguation

In some CPGs, decision variables in recommendations can be vague and not mutually exclusive. For Disambiguation, Shiffman gives an example of this involving asthma exacerbation severity in which the descriptions for different severity levels contain ambiguous terms that could overlap. Ambiguity in a CPG can be the result of limited supporting evidence or lack of consensus. While the guidelines recommendations encoded into ATHENA-CDS did not require this type of disambiguation, there were recommendations that contained decision variables that have multiple incompatible interpretations (e.g., different cutoff values for abnormal laboratory test results). However, the translation of these variables into operational concepts was categorized as a deabstraction since they did not require delineating two mutually inclusive recommendations.

2.4 Verification of Completeness

This step ensures that recommendations provide guidance for all clinical scenarios. The goal of verification is to identify gaps in the decision criteria or in the recommended actions. The deabstraction and disambiguation process
often elucidates these gaps. In the example above, the criteria for starting an aldosterone antagonist do not account for the scenario where the serum creatinine and estimated glomerular filtration rate (eGFR) have not been measured. In addition, while the criteria for starting the medication allow for values of either serum creatinine or eGFR, the subsequent dosing recommendations provide dosing recommendations based on eGFR but not serum creatinine.

2.5 Build Executable Statements

After the guideline knowledge has been atomized, deabstracted, disambiguated, and verified for completeness, it is ready for translation into full executable statements that can be encoded into the CDS logic. The example we have been building in the previous steps could result in refined logical criteria such as:

\(((\text{serum creatinine} < 2.5\text{mg/dL or eGFR rate} > 30\text{ mL}) \text{ and (absence of 20\% rise in serum creatinine from the lowest recorded value in the last 90 days) and no history of serum potassium level} > 1.5\text{mEq/L above normal}}\)

This refined statement continues to have abstract or ambiguous concepts such as “normal” potassium. As can be seen, preparing guideline knowledge is often an iterative process that continues until the logic can be fully encoded.

2.6 Explanation

In addition to building executable statements, an important aspect of preparing the guideline knowledge involves providing the user with the reasoning behind each recommendation. Healthcare providers often want to know the evidence behind a recommendation when it goes against their normal practice. Explanation can also fill gaps in provider knowledge. In addition, user interface screen real estate is often limited, preventing the display of every detail of a recommendation or the evidence.

2.7 Knowledge Representation and Encoding*

This is a new step that encompasses the process of encoding the knowledge prepared in the previous steps. Shiffman presumed that the knowledge would be encoded in rules with decision variables. We implemented ATHENA-CDS using a rich guideline model that required decisions on specifying how the knowledge will be represented (e.g., how a domain ontology that includes taxonomies of medical conditions and drug interventions should be organized and how incomplete information should qualify suggested drug recommendations). In addition to deciding how the logic will be encoded, details on standards and conventions, version control, and the organization of data are important considerations that will facilitate the implementation process as well as future maintenance and extension. This step also includes making adjustments to the CDS logic based on the CDS design specifications described below in the steps involving EHR and Local System Integration. Once the knowledge representation has been determined and the knowledge has been aligned to specifications, the knowledge encoding can be performed.

3. EHR and Local System Integration

3.1 Workflow Integration

3.1.1 Data Origin Identification

In order to understand how the CDS integrates into the workflow, it is necessary to identify the clinical data sources. Data can come from historical records stored in the data warehouse. Other data can originate from certain steps in the clinical workflow and get recorded at point of care through data entry into the CDS or EHR.

3.1.2 Recommendation Insertion Identification

Determining how users will want to use the system and where recommendations should be inserted in the workflow informs design development to enhance the adoption and effectiveness of the CDS. For example, ATHENA-CDS generates recommendations on guideline-concordant changes to a patient’s active medication regimen. As a result, the CDS was designed so that the recommendations are presented to the clinician while he/she is reviewing data at the point of care and before completing the order entry process.

3.2 CDS Alignment, Adaptation, and Enhancement

3.2.1 CDS Alignment*

In many healthcare organizations, there are multiple types of CDS integrated into the provider workflow. A successful CDS implementation ensures appropriate integration in the existing architecture. For this project,
ATHENA-CDS was being integrated within a clinical dashboard showing metrics on specific performance targets. Since decision criteria and therapeutic targets drive the recommendations provided by ATHENA-CDS, they were modified to align with the numerator and denominator criteria of the performance metrics implemented by the dashboard. This alignment enables a user to have a consistent experience across the system. CDS Alignment can also involve the process of specifying the responsibilities of a CDS when there is an overlap in functionality with other systems. Unlike the “CDS Enhancements” step (discussed below), CDS alignment does not create new CDS capabilities, but utilizes existing capabilities. In contrast to the “Associated Beneficial Services” step (also discussed below), this step involves modifications to CDS capabilities so that they are consistent with existing CDS services.

3.2.2 CDS Adaptation*

Similarly, adaptation with the local EHR must be considered in the implementation of guideline recommendations. Specific components of the recommendation might need to be aligned with data available to the CDS. For example, if clinical information is not readily retrievable from the EHR, then alternative data sources must be established or limitations of the system must be acknowledged. Early consideration of the local system in the knowledge preparation process can reduce the chance of late modifications or removal of a recommendation. For example, many veterans receive obstetrics care outside of the healthcare system and pregnancy is not always documented in the EHR. As a result, a decision was made to consider pregnancy out of scope of ATHENA-CDS, and providers were advised to use their clinical judgment in women of reproductive age. CDS adaption involves changes to the implementation, not because of existing CDS capabilities that need to be aligned, but because of the characteristics of the local population and the data available in the EHR.

3.2.3 CDS Enhancements*

During the implementation and subsequent maintenance of a CDS project, new healthcare information technology (HIT) can emerge and interoperability between systems can improve. The availability of new data and enhancements to the EHR may allow for the implementation of new CDS capabilities. During the ATHENA-CDS implementation, the ability to check a patient’s adherence to prescriptions based on medication possession ratios became available through pharmacy data analytics. This new capability allowed for the implementation of CPG recommendations that could not be implemented before. In contrast to the use of associated beneficial services to complement CDS recommendations described by Shiffman, CDS enhancements entail changes in the CDS recommendations generated by the system.

3.3 CDS Delivery

3.3.1 Action Type Definition

The culminating step in CDS is an output that is ultimately used clinically. The majority of outputs can be classified in three action categories: gathering information, interpreting information, performing a task, and organizing care. Gathering information can come in the form of monitoring a clinical variable according to a specific criteria or schedule. Interpreting information usually results an output such as a diagnosis, prognosis, or clinical status. Recommended tasks can include prescribing medications, performing a procedure, documenting in the medical record, advocating a policy or practice, or preparing for a guideline-directed activity. Organizing care involves directing the flow of care for a patient in the form of deposition, follow-up, and referral.

3.3.2 Associated Beneficial Services

Shiffman describes this step as linking action types to services that offer design patterns for facilitating clinical care. Common use cases involve linking a prescription with a pharmacy system, sending an study order to a lab, or sending a referral to the designated department. These use cases facilitate the implementations of specific CDS recommendations using new or existing services, but do not alter the CDS recommendations themselves. For this project, ATHENA-CDS was being integrated with an existing clinical dashboard with an established set of associations. No additional decision regarding beneficial services arose during the development of the system.

3.3.3 Terminology Mapping*

An increasingly common task in HIT is data mapping to facility interoperability between systems that use different naming conventions. As a result, terminologies and concepts encoded into the knowledge representation such as laboratory studies and medications have to be mapped to those in other systems. For our project, a mapping table was needed for the data retrieval interface between ATHENA-CDS and the EHR. Similarly, the output of ATHENA-CDS had to be mapped to the clinical dashboard’s data repository and terminology usage.
3.2.4 User Interface Design

Best principles of user interface design must be used to make an effective and usable CDS. In addition to the standard design issues such as formatting of the look and feel to ensure readability and usability, consideration must be made in terms of user workflow, balancing space-efficiency, and the number of clicks required to perform a task. These design decisions can be constrained since most CDS are integrated into an existing system and must conform to or stay within the capabilities of the existing design.

4. Testing

Verification and validation of the encoded system is vital before a CDS system is deployed and should be performed throughout the development process. Verification is the process of ensuring the system operates according to requirements specifications. During this process, encoded logic and knowledge are tested to make sure they work as intended. CDS verification can be achieved though appropriate selection of real or engineered patient data with comparison to a reference standard. Validation of a CDS system ensures that the system performs what the end user requires. Validation can be achieved by quantitative and qualitative studies that measure the completeness and clinical appropriateness of system-generated recommendations.

5. Project Management

Management of the project throughout the implementation process is important in ensuring a cohesive and complete software product. There are many decisions that are related to the project as a whole. These decisions can affect the previous steps by influencing the scope of the project as well as the organization and timing of tasks. Project management also facilitates appropriate communication and resource allocation and tracks progress. For this study, these administrative decisions often related to resources such as scope, human resources, and financing. Other project management decisions that were important in a CDS implementation process included decisions on organizational policies, information technology challenges, analytics support, deployment details, and quality/feature control.

Discussion

With the immense amount of medical knowledge available, CPGs provide clinicians with access to concise expert or evidence-based knowledge sources. However, there are numerous barriers that reduce the use of CPGs in clinical practice. The Shiffman model describes a process for translating CPGs into a general requirement specification for a CDS that can be developed in isolation and implemented later in any setting. The resulting specification accounts for input variables from the CPG. In contrast, our model describes the extended process of taking a CPG, converting the knowledge into a computable format, and integrating that operationalized knowledge into a context-specific live user-facing system. As a result, the design specification and implementation process required consideration of variables specific to the VISN 21 environment as well as the national VA system.

In the analysis of our implementation, the incidence of decisions demonstrates that some steps of the original Shiffman model required more discussion and thought than others. It is possible that this variation reflects the importance and complexity of some of the steps in regards to their effect on the project as a whole. Three of the steps required a great deal of team discussion. Selection of an appropriate, usable, and valid medical knowledge is a fundamental step in clinical practice, and it is not surprising that the guideline and guideline recommendation selection process was a major point of discussion for ATHENA-CDS. Similarly, the importance of usability in the adoption and effectiveness of CDS continues to be underscored, and user interface design decisions were heavily discussed for our CDS implementation. Lastly, verifying the completeness of recommendations is a complex process that necessitates the input of both clinical and knowledge engineering experts.

Conversely, some steps, such as Markup, Atomize, De-abstract, Disambiguate, Build Executable Statements, and Define Associated Beneficial Services are either not reflected or under-represented in our formally recorded decisions. This finding can be explained by three factors. First, some of Shiffman’s and our additional implementation steps are optional. For example, the need for Shiffman’s Associated Beneficial Services and our CDS Alignment and Enhancements are contingent on the implementation context. Define Associated Beneficial Services step was performed prior to our project during the implementation of the clinical dashboard. Second, while the concepts behind the Markup and Build Executable Statement steps are still essential in implementing a narrative guideline, their definitions needed to be broadened, as there is no consensus technology used for this task. Lastly, the organization of our project team, and the limitation of using recorded implementation decisions as the basis for
categorization can explain the under-representation of steps such as Atomize, De-abstract, Disambiguate. For many of the knowledge representation tasks, our experienced knowledge engineers perform these tasks without requiring significant input and are not reflected in the recorded decisions of the project.

As more CDS systems are integrated into EHRs, more attention needs to be placed on how they interact. It is not uncommon that the logic of each system is encoded by multiple different groups or vendors. As a result, the design specifications for these CDS do not take into account other systems. Alignment of CDS and reconciliation of multiple CDS in the same system are developing areas in informatics that will need continued investigation as CDS implementations increase.

Similarly, interactions between CPGs have been a known issue in both clinical practice and CDS. CPGs, in general, are brief in their discussion of comorbid conditions. In ATHENA-CDS, each chronic disease KB is derived from CPGs, and the CDS provides recommendations independent of the other KBs. The KBs were specifically encoded to reduce possible interactions produced by the system. However, additional investigation is needed to make multimorbidity support feasible in working implementations.

While we have attempted to generalize our findings, other implementation projects will likely face different challenges. As HIT continues to evolve rapidly, the methodology and complexity of each step will likely change. Continuing to define the methods and steps in translating evidence-based knowledge into CDS will help ensure successful implementations. In addition, efforts should continue to standardize formats for guidelines that facilitate their implementation both clinically and computationally.

Conclusion

We validated and extended steps in Shiffman’s approach to making guideline recommendations computable. We identified additional knowledge engineering and implementation categories needed because of multiple comorbidities and guidelines, actual system implementation, and integration with existing HIT tools.

Acknowledgements

This work was supported in part by VA Health Services Research and Development (HSR&D) grant IIR 11-071-1 (PI: Goldstein). An Advanced Fellowship in Medical Informatics that is funded by the VA Office of Academic Affiliations, HSR&D Service, and Office of Informatics and Analytics supported Dr. Tso’s time. Views expressed are those of the authors and not necessarily those of the Department of VA or other affiliated institutions.

References


27. Adrion WR, Branstad MA, Cherniavsky JC. Validation, verification, and testing of computer software. ACM COMPUT SURV 1982;14:159-92.


Automating Performance Measures and Clinical Practice Guidelines: Differences and Complementarities

1Samson W. Tu MS, 2Susana Martins MD MSc, 2Connie Oshio PhD, 2Kaeli Yuen, 2Dan Wang PhD, 2Amy Robinson PharmD, 2Michael Ashcraft MD, 1, 2Paul A. Heidenreich MD MS, 1, 2Mary K. Goldstein MD MS

1Stanford University, Stanford, CA; 2VA Palo Alto Health Care System, Palo Alto, CA

Abstract

Through close analysis of two pairs of systems that implement the automated evaluation of performance measures (PMs) and guideline-based clinical decision support (CDS), we contrast differences in their knowledge encoding and necessary changes to a CDS system that provides management recommendations for patients failing performance measures. We trace the sources of differences to the implementation environments and goals of PMs and CDS.

Introduction

Performance measures and clinical decision support (CDS) are methods to improve quality of care. Both performance measures and CDS systems rely on clinical evidence, often summarized in clinical practice guidelines (CPGs), to define the standards of care. They are inter-related but distinct. Performance measures seek to improve care by retrospectively measuring the quality of the care provided to populations of patients, while CDS focuses on prospectively providing evidence-based therapeutic recommendations and alerts that are custom-tailored for the circumstances of particular patients. CDS may include performance measurement information as feedback to health professionals; however, in this paper we will use the term CDS in the sense of providing timely information and advisories to health professionals to assist decision-making. Once performance measures have been established, health care systems can provide CDS in order to improve performance on the items that are being measured. To improve the quality of care beyond providing aggregated data, the first step is feedback about performance for each patient with respect to the targets being measured. The next step is to add CDS to give additional recommendations about how to manage the condition to achieve the target.

In this paper we seek to concretely characterize the ways performance measures and guideline-based CDS differ yet can be complementary to each other by closely analyzing two pairs of performance-measure and CDS implementations. In the first pair (Analysis 1), we examine the implementations of similar clinical recommendations as performance measures and as CDS, and highlight how the implementations differ in their workflow integration, cohort definitions, definition of compliance, use of data, and output formats. We categorize the rationales for the divergence in inclusion and exclusion criteria of performance measure and CDS. In the second pair (Analysis 2), where a CDS system is used to provide guidance on the management of patients who have failed particular performance measure, we describe the necessary changes to the implementation of CPGs in a CDS system and how the recommendations of the CDS system can complement performance measure status in a provider’s dashboard.

Background

The Veteran Health Administration (VHA) of the Department of Veterans Affairs (VA) has been a leader in health care quality assessment and improvement [1]. Not only did it implement quality-improvement efforts that were guided by performance measurements [2], it also pioneered the application of evidence-based CPGs at points of care [3]. For more than ten years, our group at VHA Palo Alto Healthcare System has used the ATHENA CDS system, a knowledge-based system to provide CDS for guideline-based care, to investigate issues related to the operationalization, testing, and deployment of CPGs [4-7]. The basic system architecture includes Protégé [8] knowledge bases (KBs) that contain computer-interpretable CPG recommendations encoded using a domain-independent guideline model, a guideline interpreter execution engine that applies the encoded recommendations to patient data to generate patient-specific recommendations, and client programs that display the recommendations to and interact with CDS users. Initially focused on hypertension, the ATHENA CDS system now includes knowledge bases in several other clinical domains, such as hyperlipidemia, chronic kidney disease (CKD), diabetes mellitus (DM), heart failure (HF), and opioid therapy for non-cancer chronic pain [9-11].

For each knowledge base, the ATHENA CDS system evaluates decision criteria to determine guideline-concordant management goals and recommended actions for a patient. This computational infrastructure of the CDS system
would seem well suited to evaluate performance measures as well, since implementing performance measures involves using similar data and criteria. Performance measures focus on numerator and denominator criteria to determine whether a patient is included in the target population of the performance measure (the denominator criteria), and, if included, whether their care satisfies the definition of quality care (the numerator criteria). In 2011, our group was afforded the opportunity to study the processes and results of automating performance measures and guideline-based recommendation for patients diagnosed with heart failure in ATHENA CDS.

In a subsequent effort, we adapted the ATHENA CDS system to provide guideline-based recommendations to improve the care of patients who fail performance measures. Within the VA, the Veterans Integrated Service Network (VISN) 21 Pharmacy Benefits Management (PBM) group has developed a clinical data warehouse, based on an SQL Server database. The PBM group has built a clinical dashboard for use by both managers and individual providers, including nurses, pharmacists, physicians, and other members of the health care teams. The dashboard provides tools to monitor the clinical performance measures used by VA, focusing on the diabetes, heart disease, and hypertension measures as the priority areas identified by the leadership. The clinical dashboard is available to clinical managers for information about the primary care providers they manage and to primary care providers for managing their own panels of patients. The dashboard provides a stoplight-type report (red/yellow/green on each performance measure) which can be viewed both as a “panel” view (a provider’s panel of patients) or as a visit view (patients coming into clinic today, for visit planning). Our group undertook a project to complement the dashboard’s implementation of performance measures with detailed CDS for patients who fail to satisfy these performance measures. This opportunity allows us to examine what is different about a CDS system used in the electronic health record vs a CDS system used within a performance measure dashboard.

Method

To investigate concretely how automated performance measure systems and guideline-based CDS systems differ from and complement each other, we performed two analyses.

Analysis 1: Comparing Implementations of Heart Failure Performance Measure and CDS Systems

The analysis consists of comparing a heart-failure performance-measure (HF-PM) system that implements National Qualify Forum (NQF) Measure 0081 on the use of angiotensin-converting enzyme (ACE) inhibitor or angiotensin receptor blocker (ARB) therapy for left ventricular systolic dysfunction [12] and the ATHENA HF-CDS system that implements similar recommendations from the 2013 ACC/AHA guideline for the management of heart failure [13] (See Figure 1).

ATHENA CDS systems structure recommendations in terms of the EON guideline model [14, 15]. This guideline model formalizes a CPG as a knowledge structure containing eligibility criteria, goals or targets of therapeutic interventions, and a clinical algorithm that provides distinct decisions and action choices for patients in various clinical scenarios. The guideline model includes expression languages for performing queries and for encoding decision criteria [14]. At run-time an expression-evaluation module of the execution engine uses patient data to evaluate expressions and conclude whether a decision criterion evaluates to true or false for a patient. The evaluation of decision criteria helps to generate therapeutic recommendations appropriate for a particular patient.

To implement performance measures, we extended the EON modeling and execution infrastructure. Because we wish to compute a collection of performance measures using the same data set, we organize performance measures into groups, such as measures applicable to inpatient cases and measures that are applicable to outpatient cases. To improve system efficiency, we identified, for each group, common criteria that are applicable to all measures in the group. (For example, for the outpatient performance measures, one required criterion is that a patient has an outpatient encounter during the measurement period.) These common criteria (implemented as EON eligibility

---

1 We actually implemented NQF performance measures 0081 and 0083, where the NQF 0083 is a measure that evaluates the use of beta blockers for patients with heart failure. For the sake of simplicity, we report the results derived from the use of ACE inhibitor and ARB. The conclusions that can be drawn from the use of beta blocker are similar.

2 Specifically, we implemented recommendations related to ACE inhibitors and ARBs in [13]: ‘ACE inhibitors are recommended in patients with HFrEF and current or prior symptoms, unless contraindicated, to reduce morbidity and mortality. (Level of Evidence: A)’ – Yancy 7.3.2.2.’ and ‘ARBs are recommended in patients with HFrEF with current or prior symptoms who are ACE inhibitor intolerant, unless contraindicated, to reduce morbidity and mortality (Level of Evidence:A) – Yancy 7.3.2.3
criteria) are part of a performance measure’s denominator criteria. Each performance measure within a group has a set of inclusion criteria (all of which must evaluate true for a patient to be in the denominator population), a set of exclusion criteria (any of which, if evaluated to true, excludes a patient from the denominator population), and a set of criteria to achieve (any of which, if evaluated to true, puts the patient in the numerator population). Because modeling a performance measure’s numerator and denominator criteria uses the well-tested EON expression language for encoding the criteria and executing them against patient data, these extensions were easily implemented.

While NQF 0081 provides an initial level of specifications regarding the numerator, denominator inclusions, and denominator exclusions for these measures, in order to operationalize the computation of the measures we had to interpret the measures in much more detail. For instance, the denominator exclusions in NQF #0081 are defined quite broadly as follows: “Documentation of medical reason(s) for not prescribing ACE inhibitor or ARB therapy”, “Documentation of patient reason(s) for not prescribing ACE inhibitor or ARB therapy”, “Documentation of system reason(s) for not prescribing ACE inhibitor or ARB therapy.” In order to better specify such broad exclusion criteria, we consulted other heart failure performance measures that had more specific definitions for denominator exclusions. For instance, the VA External Peer Review Program (EPRP) has a FY2012 technical manual that specifically defines an inpatient performance measure of “HF patients with left ventricular systolic dysfunction (LVSD) who are prescribed an ACE INHIBITOR or ARB at hospital discharge.” The measure’s denominator exclusions include: “Patients who had a left ventricular assistive device (LVAD) or heart transplant procedure during hospital stay” and “Patients who have a Length of Stay greater than 120 days”. Similarly, other NQF- endorsed heart failure performance measures related to ACE inhibitors and ARBs, such as NQF #0610, specify additional denominator exclusions, such as “Evidence of metastatic disease or active treatment of malignancy (chemotherapy or radiation therapy) in the last 6 months.” We have pooled the exclusion criteria from these various sources and used them as operationalized denominator exclusions of NQF #0081.

We evaluated the performance of HF-PM using a convenience sample of 340 VA patients. Out of the 340 patients, 73 outpatient cases and 33 hospitalizations satisfy initial eligibility criteria. A preliminary validation of the accuracy of the system on 12 inpatient hospitalizations and 20 outpatient cases demonstrates that the system successfully generates conclusions for the ACE-inhibitors/ARB and beta-blockers performance measures in the majority of cases.

To operationalize the ACE inhibitor and ARB recommendations in the 2013 ACC/AHA guideline, we followed the methodology outlined by Shiffman [16] for making explicit the translation of document-based knowledge: markup of the text; atomization; de-abstraction; disambiguation of concepts; verification of completeness; and addition of explanations. We have identified additional steps not in Shiffman’s categorization and have reported them in a separate paper [17]. In the rest of the paper, we refer to the ATHENA CDS system that implements the heart-failure CPG recommendations as HF-CDS.

Implementing similar recommendations, first as performance measures and then as part of a CDS system, afforded us the opportunity to systematically analyze how different usages of the same recommendations have implications for workflow integration, cohort definition, definition of compliance, use of data, and output formats. We describe the findings in the Results section.

Analysis 2: Use Guideline-Based CDS to Complement Performance Measure Evaluation

The second analysis involves a production implementation of performance measures in the VA VISN 21 dashboard and a modified ATHENA CDS system (called ATHENA PMtoCDS) designed to provide decision support on the management of patients who have failed the performance measures. To provide recommendations that are consistent with performance measure evaluations, we made the necessary changes to the existing CDS implementations of guidelines for the management of type II diabetes mellitus, hyperlipidemia, heart failure, chronic kidney failure, and hypertension. We also developed a prototype user interface that integrates the outputs of the performance-measure dashboard and ATHENA PMtoCDS. The user interface allows a user to drill down from the top-level display of performance measure evaluations to see recommendations from ATHENA PMtoCDS on how to improve compliance with the performance measures (Figure 2). In this pairing of performance-measure and CDS systems, we demonstrate the complementarity of the two systems in achieving the institution’s clinical objectives.
Figure 1. In the first pair of systems, the Heart Failure Performance Measure system (1a) is designed to generate reports on whether a cohort of patients satisfies the NQF 0081 performance measure. ATHENA HF CDS System (1b) is designed to generate detailed management recommendations for guideline-based care at the time of a patient encounter.

To make the CDS advisories consistent with performance measures, we modified the original encodings of the guideline recommendations. The advisories generated by ATHENA PMtoCDS for patients who fail specific performance measures are passed to the dashboard for display to users in the context of the performance measure evaluations. Parallel to Analysis 1, we report as results our analysis of the workflow integration, cohort definition, compliance definition, use of data, and design features of user interface of the CDS. In this analysis our focus is on changes that are necessary to integrate CDS recommendations in the context of a performance-measure-oriented dashboard.

Results

Table 1 at the end of section highlights the main results of the analyses.

Analysis 1: Comparing Implementations of Heart Failure Performance Measure and CDS Systems

We will contrast performance measures and guideline-based CDS implementations in terms of workflow integration, cohort definitions, compliance definition, use of data, and output structure.

1. Workflow Integration

Both the HF-PM and the HF-CDS systems are research prototypes that were not deployed in actual clinics. The design of the HF-PM system involves the system processing, in batch mode, the data of a cohort of patients to see whether their treatments satisfy performance measures. The HF-PM system evaluates performance measures for both inpatient and outpatient cases. Both aggregated results and results for individual patients are stored in a database accessible to providers or administrators when desired. The HF-CDS system, on the other hand, evaluates patient-specific current clinical care and prospectively recommends best practices based on the encoded guideline recommendations. The CDS system also brings relevant patient data into one display with layered information to reduce the cognitive burden of searching the EHR for the information required for decision-making. Past ATHENA CDS deployments focused on providing decision support to primary care providers in outpatient clinics. For example, ATHENA HTN, the hypertension management version of ATHENA CDS, would pop up a CDS window when a provider selected a patient who was eligible for guideline-based care. The window would contain recommendations based on the JNC/VA guideline for the management of hypertension and the available patient data.

2. Cohort Definitions

Next we report the results of the comparison between HF-PM and HF-CDS systems in terms of identifying patients who should be included in performance measure evaluations and CDS support. As described in the Methods section,
we use as a case study the ACE inhibitor recommendation based on NQF 0081 [12] and the 2013 HF guidelines [13]. To make the criteria comparable, we examine the cohort definition of the outpatient component of NQF 0081 only.

HF-CDS makes use of 13 criteria that determine whether or not a recommendation should be made on ACE inhibitor or angiotensin II receptor blocker (ARB) for a patient with heart failure. HF-PM uses 33 criteria to select the cohort of patients eligible to have a prescription for an ACE inhibitor or an ARB.

We identified 8 criteria that are identical, 16 HF-PM criteria that are handled differently in HF-CDS, 10 criteria that are in HF-PM only and 3 criteria that are in HF-CDS only.

• Similar criteria differed in definition and modeling choices
  We organize the 16 instances in which similar criteria were handled differently in HF-PM and HF-CDS systems into 3 categories:
  a. Role of clinical judgment: HF-CDS alerts providers to conditions that require clinical judgment when recommending ACE inhibitor while HF-PM excludes these patients from the cohort/denominator in order to improve specificity. Performance measures exclude specific reported adverse events such as hypotension, hyperkalemia, worsening renal function due to ACE inhibitor or ARB. HF-CDS displays both the recommended drugs and the adverse events and allergies to providers and leaves the choice of whether to prescribe the recommended drugs to their judgment. Other clinical conditions were also excluded by performance measure, such as presence of aortic stenosis, hypertrophic cardiomyopathy, renal artery stenosis, stage 3 chronic kidney disease, eGFR between 30 and 59, and active prescription for aliskiren, are handled in the HF-CDS by explicitly alerting the provider to these conditions.
  b. Differences in sources: First, the heart-failure guideline used as the basis for HF-CDS does not contain an enumeration of ICD 9 codes used to define heart failure. Subject matter experts weighed in to define the list of ICD 9 codes and included cardiomyopathy codes that were excluded in the sources that we used to define heart failure for NQF 0081. Second, the definition of the thresholds for ejection fraction in the performance measure was more stringent (less than 40) than the threshold used for HF-CDS (less than or equal to 40).
  c. Differences in the timing and retrospective/prospective nature of performance measure and CDS: Pregnancy was modeled differently in the two systems. In HF-PM we look at the data retrospectively for pregnancy codes in order to exclude pregnant women from the denominator population. HF-CDS was designed to give close to real time advice. Given that many VA patients receive pregnancy care outside of VA and thus the relevant pregnancy codes may not be up-to-date in VA data, we decided to issue an alert to all women of childbearing age about the use of ACE inhibitor.

• Criteria unique to HF-PM
  HF-PM had 10 unique criteria that are not applied in HF-CDS. They can be grouped into two categories:
  a. Active in health care system: HF-PM excludes patients who are not active in the health care system using criteria such as visit in past 12 months and absence of death. In HF-CDS there is no need for this filter since CDS applies to patients with a scheduled visit, and are already known to be active in the health care system.
  b. Differences in sources: Our modeling of performance measure was based on documentation from national sources that mentioned exclusion criteria that were not cited in HF guidelines. These include use of hydralazine prior to ACE inhibitor/ARB (NQF 0610), multiple myeloma (NQF 0610), active prescription for pulmonary hypertension medications (NQF 0610), heart valve surgery (NQF 0610) and previous admission for hyperkalemia (NQF610).
  c. Differences in goals: HF-PM excludes patients with potential limited life expectancy while the HF-CDS offers the recommendations since the life-prolonging measures are often also measures that reduce symptoms and contribute to quality of life.

3 The categorized HF-PM criteria sum to 34 because one HF-PM criterion, eGFR<60, is broken up into eGFR<30 and 30<=eGFR<60, as HF-CDS uses them differently. The categorized HF-CDS criteria do not sum to 13 because multiple HF-PM criteria (e.g., adverse reactions) are handled uniformly in HF-CDS (e.g., adverse reactions not treated as absolute contraindications are displayed with recommended drugs).
• Criteria unique to HF-CDS

The three criteria that are unique to HF-CDS-only can be classified into three types:

a. Differences in sources: HF-CDS used creatinine values in addition to eGFR and ICD 9 codes for chronic kidney disease to exclude patients.

b. Difference in the scope of CDS: Because of the complexity of managing end-stage heart failure, stage D patients are excluded from the scope of HF-CDS.

c. Differences in output: Because HF-CDS generates recommendations on the choice of medications to prescribe, the presence of ACE inhibitor or ARB on the current medication list suppresses such recommendations.

3. Definition of Compliance

The NQF 081 performance measure defines a patient satisfying the denominator criteria as compliant with the performance measure if the patient was prescribed either ACE inhibitor or ARB. HF-CDS, on the other hand, preferentially recommends ACE inhibitor as the drug of choice, because of its well-founded evidence base and lower cost, and recommends ARB only if a patient is intolerant of ACE inhibitor.

4. Use of Data

Although both systems have access to all data in the EHR we note a distinction in temporal requirements when defining eligible patients. ATHENA CDS uses all data available in the electronic medical record with limited temporal restrictions on specific criteria while performance measure applies stricter temporal limits in many inclusion and exclusion criteria. For example, the NQF 081 performance measure requires the presence of a heart failure ICD9 code in the 2 years prior to the measurement period while for HF-CDS it is sufficient to have an ICD9 code for heart failure at any time. This difference supports the goals of the distinct systems: in performance measure to improve the specificity of the cohort leading to better credibility of results and in CDS to provide full presentation for review by providers in the context of care.

5. Output Formats

The performance measure primary output is geared to administrators wishing to evaluate the quality of care provided to patients at a given site or over a specific region or to health care professionals to identify missed opportunities to improve care. For HF performance measure, the raw output for each patient consists of a listing of each inclusion, exclusion, and numerator criteria and the evaluated results as determined from patient data. This listing is then transformed into summary statistics showing the number of patients who qualified for the measure and what proportion of these met the performance measure (i.e. met the numerator criterion). Secondary analyses showing various percentages of patients meeting certain exclusion criteria or failing to meet certain inclusion criteria are also computed to facilitate understanding of various contributing factors.

In contrast, the CDS output is a patient-specific advisory geared towards providers with the assumption that they will review the information and apply their clinical judgment. It is in a layered graphical user interface. The top level contains patient data, whether a patient reached the guideline goal, alerts, and detailed therapeutic recommendations. Therapeutic recommendations contain multiple guideline-compliant choices, each with patient specific indications, contraindications and adverse events highlighted. In a second layer we provide additional drug-related information such as the need for monitoring, drug dosing and other relevant alerts. The objective of the design is to bring the most important information for the decision making process upfront and to display additional relevant information upon user demand.

Analysis 2: Use Guideline-Based CDS to Complement Performance Measure Evaluation

To assist clinicians providing best practices in the performance measure environment we modified ATHENA CDS, originally developed as a standalone CDS system, into ATHENA PMtoCDS, a CDS system that provides advisories for patients failing performance measures. In the following we describe the changes made so that ATHENA PMtoCDS can play this role.

1. Workflow Integration

Instead of triggering CDS at the point of care, ATHENA PMtoCDS is designed to display pre-computed CDS recommendations when a provider reviews a panel of patients on the VISN dashboard. The software that computes dashboard performance measure is configured to generate a set of cases that fail the performance measures and to
pass the set to ATHENA PMtoCDS, which then generate the CDS. The dashboard displays patients’ performance with respect to VA performance measures as a population-based summary and as individual patient records. For a patient who fails to meet selected performance measures, a user can bring up ATHENA PMtoCDS recommendations that suggest how to manage the treatment of that patient.

2. Cohort Definitions

Because the performance measure system filters out patients who satisfy performance measure targets, CDS recommendations will not be generated for those patients. As originally implemented, ATHENA CDS provides recommendations regarding all patients independent of their treatment goal. For example, in the ATHENA-Hypertension CDS system, if a patient’s blood pressure is within the target range the system may recommend substitution from less desired to more preferred antihypertensive agents. These recommendations would never be generated in the performance measure environment since only patients with blood pressures above target would be eligible for the CDS. Another example is the way CDS manages glycemic control in patients with type II diabetes. The VA performance measure for the target HbA1c is less than 9% while in a point-of-care CDS system the target would be set by the provider in conjunction with the patient and can change over time. The cohort of patients receiving ATHENA PMtoCDS recommendations is restricted to those whose HbA1c is greater than 9%.

3. Definition of Compliance

As described above, having ATHENA PMtoCDS work in conjunction with the performance measure dashboard means changing the HbA1c goal from an individualized, patient-provider agreed-upon target HbA1c to a single HbA1c goal (of less than 9%). Because recommendations and messages are given only to those who fail this performance measure, and are above this threshold, there are recommendations that would either not be given, or given only regarding a subset of patients rather than all patients. For example, if a patient has a HbA1c<6%, and meets certain other criteria the CDS would issue a message that he/she is potentially at increased risk for cardiovascular events; this message would not be issued when considering only those patients with a high HbA1c. Similarly, if a patient’s bicarbonate level is less than or equal to 21 mEq/L, an alert would normally be given regardless of their HbA1c level. Thus because of the pre-filtering of cohort receiving CDS, some guideline recommendations or alerts become inapplicable or are applied to a subset of patients.

4. Use of Data

The differences described in the results of Analysis 1 in how performance measures and CDS use patient data also apply to Analysis 2. Integration with the dashboard, however, enables the CDS to access data that are computed by the dashboard. This allowed for the implementation of CPG recommendations that could not be implemented before. For example, the dashboard computes the medication possession ratio (MPR) that indicates the level of adherence to the prescribed medications. If the MPR is less than 0.9, then ATHENA PMtoCDS issues a message, alerting the provider that the patient may not to be adhering to his/her therapy and re-evaluation of clinical strategy may be warranted. A recommendation to increase dose or to add additional medications may not be appropriate if a patient is not adhering to the prescribed regimen of existing medications.

5. Output Formats

Unlike the previous versions of ATHENA CDS, which were directly integrated with the VA’s CPRS, the display is generated using the same Microsoft Report Server tool that generates the dashboard performance results. The advantage of using the same Report Server tool for display is that consistency of text font, size, colors, and navigation icons, i.e. the look and feel, between the CDS display and other dashboard pages can more easily be attained. Even then, to minimize possible confusion it is necessary to ensure that displayed terms from the ATHENA PMtoCDS output are made consistent with their counterparts in the dashboard.

The findings of the two analyses are summarized in Table 1. The ‘Analysis 1’ columns contrast the HF-PM and HF-CDS applications along the dimensions of workflow, cohort definition, definition of compliance, use of data, and output format. The ‘Analysis 2’ column indicates the changes to the CDS system so that it supplements a dashboard application that computes and displays performance measure information. For those patients who fail to satisfy specific performance measures, ATHENA PMtoCDS recommends changes to the management of the patients according to clinical practice guidelines relevant to the performance measures.
Table 1. Major findings: (1) Comparison of HF-PM and HF-CDS systems (Analysis 1) and (2) Necessary changes to ATHENA PMtoCDS for it to supplement a dashboard application that computes and displays performance measure information (Analysis 2).

<table>
<thead>
<tr>
<th>Analysis</th>
<th>HF-PM</th>
<th>HF-CDS</th>
<th>Necessary Changes to PMtoCDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow</td>
<td>Retrospective/on demand</td>
<td>Prospective/event driven</td>
<td>Triggered as part of PM Dashboard for patients failing PM</td>
</tr>
<tr>
<td>Cohort Definition</td>
<td>More restrictive to improve specificity; retrospective data more available</td>
<td>Less restrictive to allow clinical judgment; Data less available; Exclude patients who do not need recommendation; Exclude cases requiring complex management</td>
<td>Limited to patients failing PM</td>
</tr>
<tr>
<td>Definition of Compliance</td>
<td>Single numerator metric</td>
<td>More nuanced definition of best practice</td>
<td>Modified to be consistent with PM Dashboard; No recommendations for those who pass PMs</td>
</tr>
<tr>
<td>Use of Data</td>
<td>Stricter temporal limits to improve specificity</td>
<td>More inclusive</td>
<td>Able to use Dashboard analytics (medication possession ratio) to enhance CDS</td>
</tr>
<tr>
<td>Output Format</td>
<td>Population-based summary statistics</td>
<td>Patient-specific</td>
<td>Made to be consistent with that of PM Dashboard</td>
</tr>
</tbody>
</table>

Discussion

We use the implementations of two pairs of performance measure and CDS systems to demonstrate concretely how evaluating performance measures and CDS differ and yet can complement each other. In the first analysis, we compare the implementations of similar performance measures and guideline recommendations in the same computing environment, and systematically analyze how the resulting systems differ in workflow, definition of cohort, definition of compliance, use of data, and output formats. We attribute the differences in cohort definitions to discrepancies in the sources of the performance measures, the retrospective/prospective differences of performance measures and CDS, and, above all, in the role of clinical judgment in providing CDS. While performance measures need to be applied more narrowly to a denominator population for which the numerator criteria unambiguously apply, CDS recommendations can be provided to target patients whose management is entrusted to providers who can exercise their judgment based on data and knowledge that may not be available to the CDS system.

Analysis 2 explores the necessary changes to a CDS system for it to complement the performance measure services provided by the dashboard. We see how the pre-filtering of patient eligibility limits the recommendations that a CDS system can provide and how the consistency requirement between the performance measure system and CDS system forced the CDS system to re-define the therapeutic targets to achieve and to modify its output format.

In Analysis 1 we repurposed the ATHENA CDS infrastructure to compute performance measures. Even though CDS and performance measure program appear, on the surface, to be quite different because they produce different outputs, both programs are based on the recommendations from the same clinical evidence. Seen in this way, the underlying building blocks for both are either the same or very similar. Both make use the same medical conditions, medication lists, ICD9 codes, laboratory measurements and numeric cutoff criteria. Because we had previously encoded the heart-failure recommendations in a Protégé knowledge base, these building blocks, in the forms of Protégé classes, formalized criteria, and mappings from VA data sources, were already available for reuse when we encode performance measures for the heart-failure performance measure project. Having control of the execution engine also allows us to annotate the computed results with explanations (e.g., specific patient data causing a patient to fail an inclusion criterion) that help a user understand the computed performance measures. The disadvantage of repurposing the CDS infrastructure is that ATHENA CDS is designed to provide decision support for individual patients. It suffers from performance issues when computation is required for a very large patient cohort. Using SQL to implement performance measures, as the VISN 21 dashboard does, means that queries are applied to sets of patient data at a time, making the computation much more efficient.

Placing CDS in the service of improving the achievement of performance measures is not the only way to relate CDS and performance measures. In the literature we see additional possibilities. Fonarow et al.[18] used performance measures to measure the effectiveness of guideline-based CDS. They showed improvements in 5 out of...
7 process performance measures among HF patients after 24 months of intervention that included clinical decision support tools, structured improvement strategies, and chart audits with feedback. LeBresh et al. [19] developed a “Get With The Guidelines” program that integrated a patient management tool with performance measures and guideline summaries. They found significant improvements from baseline to the fourth quarter in 11 of 13 measures[19]. Neither the Fonarow et al. nor LeBresh et al. studies used CDS that provides patient-specific management recommendations generated from automated CPG. Walter at al [20] used the best practices defined in CPGs to create performance measure. Using colorectal cancer screening as a case study, they identified a number of pitfalls in using CPGs to define performance measures that are similar to our findings about the differences between CDS and performance measures, such as not accounting for provider judgment when scoring performance measures. Finally, van Gendt et al. [21] used performance measures developed independently of CPGs to evaluate completeness of CPGs and to offer the possibility of improving CPGs. The authors formalized performance measures as goals that CPG recommendations should satisfy. For example, a performance measure may measure the percentage of diabetes patients with albumin value measured in a 12-month period. They analyzed a Type-2 DM CPG to see whether the paths in the CPG would achieve the goal of having an albumin measurement within that period. They found that, out of 35 performance measures studied, 25 (71%) suggested that there are problems with CPGs they used in the study.

Conclusion

Our analyses demonstrate that (1) the same evidence-based treatment recommendations are implemented differently as performance measures and as management advice in a CDS system and (2) a CDS system can help patients achieving compliance with performance measures but would require significant modifications so that it complements a performance dashboard in a consistent manner.

Acknowledgement

The work was supported by VA HSR&D grant IIR 11-071 and VA HSR&D Quality Enhancement Research Initiative (QUERI) grants RRP 11-428 and RRP 12-447. Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or other affiliated institutions.

References


Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record

Yan Wang, PhD\(^1\), Elizabeth S. Chen, PhD\(^4\), Serguei Pakhomov, PhD\(^{1,2}\), Elizabeth Lindemann, BS\(^1\), Genevieve B. Melton, MD, PhD\(^{1,3}\)

\(^1\)Institute for Health Informatics, \(^2\)College of Pharmacy, and \(^3\)Department of Surgery, University of Minnesota, Minneapolis, MN;

\(^4\)Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

The electronic health record (EHR) provides an opportunity for improved use of clinical documentation including leveraging tobacco use information by clinicians and researchers. In this study, we investigated the content, consistency, and completeness of tobacco use data from structured and unstructured sources in the EHR. A natural language process (NLP) pipeline was utilized to extract details about tobacco use from clinical notes and free-text tobacco use comments within the social history module of an EHR system. We analyzed the consistency of tobacco use information within clinical notes, comments, and available structured fields for tobacco use. Our results indicate that structured fields for tobacco use alone may not be able to provide complete tobacco use information. While there was better consistency for some elements (e.g., status and type), inconsistencies were found particularly for temporal information. Further work is needed to improve tobacco use information integration from different parts of the EHR.

Introduction

Social and behavioral factors such as tobacco, alcohol, and drug use are increasingly recognized as key factors for many causes of disease, disability, and mortality in the United States. A number of studies have been published describing the linkage between behavioral risk factors and their associated morbidity or mortality\(^1\)\(^4\). For example, worldwide, direct tobacco use is responsible for more than 5 million deaths each year\(^5\). The National Academy of Medicine (NAM; formerly Institute of Medicine) have emphasized the need for improving existing datasets, developing new data sources, and establishing strategies and models for incorporating social and behavioral factors and their interactions in its 2006 report on “Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate”\(^6\). In a recent NAM report on “Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2”, tobacco use and exposure was featured among the domains recommended for inclusion in the electronic health record (EHR)\(^7\).

The widespread adoption of EHR systems, in turn, provides an opportunity for clinicians and researchers to access a large amount of information about an individual’s social history including substance use. Within EHR systems, documentation of social history, including tobacco use, can range from structured and coded data to free-text narrative. A number of studies have focused on the examination and representation of social history information\(^8\)\(^-\)\(^12\). Other studies have involved developing natural language processing (NLP) techniques for the automated identification and extraction of substance use information, with a particular emphasis on tobacco use\(^8\)\(^,\)\(^13\)-\(^18\).

In this study, we sought to investigate tobacco use information collected in multiple structured and unstructured sources within an EHR system. NLP techniques, as previously described, were used to extract tobacco use statements from free-text for comparison with structured sources in order to characterize content, consistency, and completeness of this type of information for patients within a single system.

Background

In early work, a multi-institutional study was conducted to characterize social history information in clinical notes from different sources (MTSamples\(^19\), University of Vermont Medical Center [UVMCC; formerly Fletcher Allen Health Care], and University of Minnesota-affiliated Fairview Health Services [FHS])\(^10\). We evaluated adequacy of several existing models including HL7 CDA-based models\(^20\) and openEHR\(^21\) archetypes for representing social history information. From this, initial models for tobacco, alcohol, and drug use were developed in this study. Table 1 illustrates part of the model used to represent information within a social history statement for tobacco use.

In a 2014 follow-up study, 525 tobacco use entries from the social history module of the Epic EHR at UVMCC, including structured fields (e.g., for smoking status, type, and frequency) and a free-text comment field, were
manually reviewed to characterize the contents and quality issues of the free-text comments. Results from the study showed a range of potential data quality issues between the structured fields and free-text comments.

To exploit the information from clinical text, we developed a NLP pipeline for detecting substance use (alcohol use, drug use and tobacco use) statements and extracting relevant elements of substance use. Unlike many of the previous NLP studies included in the 2006 i2b2 challenge which focused on the extraction of smoking status, our goal was to extract additional semantics related to tobacco use including tobacco use beyond smoking (e.g., smokeless tobacco) and smoking status (e.g., pack-use and temporal information). In this work, we leveraged existing linguistic resources and domain knowledge from the earlier studies, as well as the Propbank resource and the MiPACQ corpus to boost extraction performance for tobacco use elements. The resulting NLP tool achieved good performance for extracting a wide breadth of substance use free text information.

Table 1. Elements and values for tobacco use statement type.

<table>
<thead>
<tr>
<th>Tobacco use elements</th>
<th>Example value or pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>current, past, quit</td>
</tr>
<tr>
<td>Temporal</td>
<td>[in/since/until] &lt;date&gt;</td>
</tr>
<tr>
<td>Method</td>
<td>chew, use</td>
</tr>
<tr>
<td>Type</td>
<td>cigars, tobacco</td>
</tr>
<tr>
<td>Amount</td>
<td>1 pack per day, &lt;#&gt; ppd</td>
</tr>
<tr>
<td>Frequency</td>
<td>occasionally, daily, socially</td>
</tr>
</tbody>
</table>

Method

Setting and Study Design

This study involved a retrospective analysis of tobacco use information collected from clinical notes and the social history module of an enterprise implementation of the Epic EHR (Epic Systems Corporation, Verona, WI) at University of Minnesota-affiliated Fairview Health Services (FHS). Fairview Health Services had been using EpicCare in one of its physician practice groups (Fairview Physicians) for over eight years. Other practices were supported by Allscripts, Eclipsys SCM, McKesson Paragon, and paper processes. Starting October 2010, two paper-based Fairview regional hospitals successfully went live with Epic clinical and revenue cycle applications. Then in March 2011, Fairview’s two largest hospitals which are academically-based went live simultaneously.

In the social history module of Epic EHR, each entry includes structured fields for tobacco use information such as smoking status, start date, quit date, as well as a free-text entry for comments as shown in Figure 1. While some clinicians use templates that pull in information from the social history module, many clinicians continue to document tobacco use information within clinical notes in free text format outside of the tobacco use module most often as part of a social history section in the note (e.g., “The patient continued to smoke about half pack a day.”). Most of the structured social history data is entered by nurses and medical assistants during ambulatory patient visits based on direct answers from patients or on questionnaires completed by patients. Tobacco use information within a note is usually entered during patient visits and is considered a required element for face-to-face ambulatory encounters in our healthcare system (as well as most others).

Figure 1. Tobacco use data entry in the social history module including comments. ©2016 Epic Systems Corporation. Used with permission.
From 337,506 adult patients (age>=18) who had a most recent social history entry in 2015 and social history entries associated with at least two prior encounters, a set of 384 patients was randomly selected as the cohort for this study to provide for a confidence level of 95%, and margin of error of 5%. For each patient, all tobacco use entries (including structured fields and free-text comments) from the social history module and provider-authored clinical notes (e.g., progress notes, history and physical examinations, discharge summaries, and admission note) were obtained. Figure 2 shows the overall high-level process of the study.

All clinical notes used in this study were collected from the University of Minnesota research clinical data repository. The repository contains documents between 1993 and 2016 from the Epic EHR, as well as documents from affiliate clinics for variable time periods. Extracted notes were pre-processed to add newlines into appropriate places based on text features such as letter capitalization, punctuation, and special letters. Sections (e.g., “Assessment & plan” section in progress note) along with the section header (e.g., “Present illness”, “Social history” and “Assessment & plan”) within each note were extracted by an NLP section extraction component. Afterwards, text within each section was split into sentences by a sentence splitter component. The approach used in our earlier study22 was used for processing sentence content as follows. First, a classifier was used to detect tobacco use statements from sentences. Tobacco use statements were then parsed by Stanford parser26 to obtain constituent and dependency parses of each sentence. The constituent parse of a sentence provides syntactic cues for later tobacco use elements extraction while the dependency parse of a sentence provided the dependency structure of the sentence. A dependency structure represents a directed graph between the tokens of a sentence (e.g., subject, modifier and preposition). With constituent and dependency parses, tobacco use elements (“Amount”, “Type”, “Status”, “Temporal”, “Method” and “Frequency”) were extracted by a substance use element extraction tool leveraging a large vocabulary of smoking-related items.

To extract the same tobacco use elements (e.g., “Amount” and “Status”) from free-text tobacco comments in the tobacco use module (Figure 1), our approach was modified slightly. Values were first collected from structured fields in social history module of Epic EHR system and were mapped to appropriate elements (e.g., “Amount” and “Status). The resulting tobacco use information from structured tobacco use fields, free-text tobacco use comments, and clinical notes with tobacco use statements were the analyzed for content and quality issues within the cohort.

![Figure 2](image_url)

**Figure 2.** Overall high-level study process resulting in information from free-text comments, structured tobacco use fields, and clinical notes.

**Analysis of Tobacco Use Information**
The content of tobacco use elements collected from the structured tobacco use fields, free-text tobacco use comments and extracted from clinical notes with tobacco use were characterized as distributions of patients, tobacco use elements and new tobacco use elements patterns of all three tobacco use information sources. A similar coding schema developed and applied in a previous study was used to examine the consistency between structured data and information extracted from unstructured data as shown in Table 2.

As observed in the previous study, a wide range of data consistency issues exists with the tobacco use fields from the social history module and free-text comments. For example, patients may have conflicting information even within a short time span. One patient with respect to smoking status was recorded as “Passive Smoker” in the structured smoking status field while the clinical note from the same day states “trying to quit smoking” and a comment states “quit date 10/30/2009”.

Table 2. Examples of observed inconsistencies and discrepancies in tobacco use data at the individual patient level.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
</table>
| Smoking status    | Tobacco use statement and tobacco use comment inconsistent with smoking status field | Smoking status: Passive Smoker  
Comment: quit date M1/D1/Y1  
Statement: The patient is trying to quit smoking |
| Packs/day         | Tobacco use statement and tobacco use comment inconsistent with Packs/day field | Packs/day:  
Comment: half a pack a day  
Statement: He currently smokes around 1/4 pack of cigarettes per day |
| Years             | Tobacco use statement and tobacco use comment inconsistent with Years of smoking field | Years:  
Comment: chewed x 12 years, 2 tins per day  
Statement: She smokes about a pack a day for about 12 years |
| Type              | Tobacco use statement and tobacco use comment inconsistent with Type field   | Type: Cigarettes  
Comment: Occasional cigar  
Statement: Significant for tobacco use |
| Start or quit date| Tobacco use statement and tobacco use comment inconsistent with Start or Quit date field | Quit date: M2-D2-Y2 00:00:00  
Comment: quit 6 1/2 yrs ago  
Statement: The patient is a former smoker who quit on M3/D3/Y3 |

To compare tobacco use data from different sources, tobacco use elements extracted from clinical notes and comments were mapped to structured fields. The “Amount” element extracted from tobacco use statements includes broader information than “pack/year” field. For example, the value for amount in the tobacco use statement “Significant for tobacco use” is “Significant”. Also, “Temporal” element can hold more information than “start/quit date” and “years”, such as the statement “The patient is a life long nonsmoker” does not provide start or quit date.

In addition, tobacco use elements extracted in free-text comments or statements can be expressed in numerous ways (e.g., “10/02/79”, “18 Aug 2008”, “x 10 yrs”, “1 ppd” and “half a pack”). In this study, we developed mapping strategies to normalize numbers and map different expressions of “Amount”, “Packs/year”, “Start/Quit date”, “Years” and “Type” in free-text tobacco use comments and statements into the structured tobacco use fields.
Results

For the 384 patients, 377 (98.2%) had tobacco use entries with the smoking status specified as a value other than “Not Assessed” (e.g., “Former smoker” and “Never smoker”); 201 (52.3%) had clinical notes with tobacco use statements, and 68 (17.7%) of patients have tobacco use comments. Figure 4 shows the overlap of the presence of notes, comments and tobacco use entries at the patient level. As shown in Table 3, only 17.7% of the patients had tobacco use comments data; fewer patients had both clinical notes with tobacco use information and tobacco use comments; and a large portion of patients had clinical notes with tobacco use information. Figure 5 shows the distribution of elements within each of the three sources.

The numbers of elements extracted from each source are listed in Table 3. We observed that a large amount of duplicate tobacco use information exists within the structured tobacco use entries. Of 5,754 entries, only 944 (16.4%) of them were unique data with different dates since information in the social history module for patients appears often to remain unchanged across encounters and therefore the same information is propagated between encounters. For clinical notes, 590 (65.8%) out of 896 tobacco use statements were unique, and only 153 (7.9%) of tobacco use comments were unique (i.e., again due to propagation of the entry between encounters).

Table 3. Tobacco use information in different sources.

<table>
<thead>
<tr>
<th></th>
<th>Total patients</th>
<th>Patients with notes</th>
<th>Patients with comments</th>
<th>Patients with structured tobacco use fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total patients</td>
<td>384</td>
<td>201 (52.3%)</td>
<td>68 (17.7%)</td>
<td>377 (98.2%)</td>
</tr>
<tr>
<td>Patients with notes</td>
<td></td>
<td>Total statements</td>
<td>Total comments</td>
<td>Total entries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>896</td>
<td>1,937</td>
<td>11,098</td>
</tr>
<tr>
<td>Patients with comments</td>
<td></td>
<td>Total statements</td>
<td>Total comments</td>
<td>Total entries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>896</td>
<td>1,937</td>
<td>11,098</td>
</tr>
<tr>
<td>Patients with structured tobacco use fields</td>
<td></td>
<td>Total entries</td>
<td>Total entries</td>
<td>Total entries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,754</td>
<td>11,098</td>
<td>Total entries</td>
</tr>
</tbody>
</table>
Figure 5. Distribution of tobacco use elements from each source.

Figure 6 shows the number of tobacco use elements and the date entered or documented for all patients over a 14-year period (2002 to 2016). Green data points signify free-text tobacco use comments and show a similar pattern as data points from structured tobacco use fields. This is likely related to free-text tobacco use comments being entered within the same user interface as the structured data fields. Compared with the other two sources, tobacco use data points from clinical notes did not show as much of an increase with time.

Figure 6. Tobacco use elements input patterns of all patients.

Figure 7a shows longitudinal distinct tobacco use information each year in an example patient start 2004 -14 with and without tobacco use elements extracted from free-text tobacco use comments and clinical notes. Each data point represents the number of distinct tobacco use elements documented within the year. The number of unique tobacco use elements for this patient increases with inclusion of information from the unstructured data sources. New tobacco use information can be plotted for the same patient with time (Figure 7b) resulting in a similar pattern. Figure 7c summarizes new tobacco use information aggregated for the patient cohort. As shown in the three figures, tobacco use information extracted from the structured tobacco use fields combined with information extracted from tobacco use comments and clinical notes provide more complete information than data from a single data source.
Figure 7a. Distinct tobacco use elements of an example patient.

Figure 7b. Tobacco use new information patterns of the same example patient.

Figure 7c. New tobacco use elements patterns of all patients.

With respect to the quality and consistency of tobacco use information from the three sources, Figure 8 shows the consistency rate of five fields between tobacco use elements collected from the structured fields, comments, and clinical notes. Compared with other data fields, there was a large amount of inconsistency with “start/quit date”
between EHR module entries, comments, and notes. Other data fields demonstrated much better consistency, which implicates the overall reliability of tobacco use information.

**Figure 8.** Data consistency between structured data and unstructured data.

Some inconsistencies were also found within structured tobacco use fields. Below examples show part of the pipe-delimited entries of the structured tobacco use fields limited to status information, in a format of “date |tobacco_use|quit_date|pack/year|years|smokeless_tobacco”. These examples also show the large amount of duplicated tobacco use within the structured tobacco use fields.

<table>
<thead>
<tr>
<th>Date</th>
<th>Status</th>
<th>Tobacco Use</th>
<th>Quit Date</th>
<th>Pack/Year</th>
<th>Years</th>
<th>Smokeless Tobacco</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1-M1-Y1</td>
<td>Former Smoker</td>
<td>Former Smoker</td>
<td>N</td>
<td>N</td>
<td>.5</td>
<td>Never Used</td>
</tr>
<tr>
<td>D2-M1-Y2</td>
<td>Former Smoker</td>
<td>Never Smoker</td>
<td>N</td>
<td>Y</td>
<td>30.0</td>
<td>Never Used</td>
</tr>
<tr>
<td>D3-M2-Y3</td>
<td>Current Every Day Smoker</td>
<td>Never Smoker</td>
<td>Y</td>
<td>N</td>
<td>.5</td>
<td>Never Used</td>
</tr>
<tr>
<td>D4-M3-Y4</td>
<td>Current Every Day Smoker</td>
<td>Never Smoker</td>
<td>N</td>
<td>N</td>
<td>.5</td>
<td>Never Used</td>
</tr>
<tr>
<td>D5-M4-Y5</td>
<td>Current Every Day Smoker</td>
<td>Never Smoker</td>
<td>Y</td>
<td>N</td>
<td>30.0</td>
<td>Never Used</td>
</tr>
<tr>
<td>D6-M5-Y6</td>
<td>Current Every Day Smoker</td>
<td>Never Smoker</td>
<td>N</td>
<td>N</td>
<td>.5</td>
<td>Never Used</td>
</tr>
<tr>
<td>D7-M6-Y7</td>
<td>Current Every Day Smoker</td>
<td>Never Smoker</td>
<td>Y</td>
<td>N</td>
<td>.5</td>
<td>Never Used</td>
</tr>
</tbody>
</table>

**Discussion**

The results in this study provide good insight into the characteristics of tobacco use content and quality issues from structured tobacco use fields in social history module, tobacco use comments in the same module and tobacco use statement in clinical notes. This work is an initial step to extract and integrate large amount of substance use information stored within EHR systems using automatic tools based NLP techniques. The approach used in this study can be easily generalized for other substance use types such as alcohol use and drug use and potentially more broadly to data represented in both text and structure fields over time. The results illustrate some of the fundamental issues of integrating substance use information from different sources. The data used in this study is limited to a single health care system of one institution with a medium sized cohort. The results will, therefore need to be
validated in other settings. Also, the automated methods will likely need modifications before being applied to other EHR systems.

We observed data consistency issues between tobacco use information from different sources, which particularly presents challenges in integrating tobacco use information from different sources and using the combined date for decision support, research, public health, and other primary and secondary uses. While we expect for data to potentially change with time as smoking status changes (e.g., a smoker who subsequently quit), we encountered discrepancies on the same date and changes in status over time that were nonsensical. For example, the smoke status of a patient as follows changed between structured tobacco use records within structured tobacco use fields and free-text tobacco use statements. The structured tobacco use fields also failed to capture this status change. The change was documented into a clinical notes at some point in between. More efforts are needed to determine adequate strategies for addressing and integrating inconsistent data from different sources, which may also need to be based off of use cases for how the data will subsequently be used.

Besides inconsistency between tobacco use information from different sources, we also noticed inconsistencies within the structure tobacco used fields such as unchanged “Years” of tobacco use or slightly different amount for “Packs/day” (e.g., 0.5 vs. 0.3). These issues indicate that the tobacco use information extracted from structured need to be further processed and normalized before integrating with tobacco use information from clinical notes and comments. Also, these issues may indicate the need for better EHR user training or improved user interface to help user entering valid data values.

Issues were also observed with comparing certain data such as “0.5 pack” vs. “10 cigarettes” or slightly different amounts “0.75 pack” vs. “0.5 pack”, “6 years ago” vs. “6.5 years ago”. Moreover, in tobacco use statement from notes and comments, tobacco use elements can be documented using various expressions with different levels of detail (e.g. “10/02/90” vs. “Feb 90” and “half pack per day” vs. “0.5 pack/day”). These issues highlight the requirement for good mapping (e.g., “former smoker” to “quit”) and normalization (e.g., “half” to “0.5”) strategies (e.g., “0.5 pack” is same as “0.75 pack”) for NLP components.

As shown in the Figure 6, more tobacco use data were entered into the social history module in recent years. We speculate that the increased entry of data in the social history module is due to a combination of factors, likely including increased attention to entry of social history information and the requirement of asking this by certain regulations (i.e., Meaningful Use requirements).

This study is limited in one substance use type, and a next step will therefore include more substance use types such as alcohol use, drug use, or caffeine. We also noticed large amount of free-text comments and statements with smoke exposure information and use of e-cigarettes, for which we did not address in this study and which represents an area of further development. Also, except for the structured tobacco use fields in the social history module, other potential EHR fields that may also include tobacco use information were not included in our analysis.

Conclusion

Overall, we applied an NLP pipeline with components to detect tobacco use information from tobacco use statements in clinical notes and tobacco use comments from social history module of Epic EHR, as well as retrieved structured information from the tobacco use EHR module. The extracted tobacco use information was analyzed over a cohort of patients, characterizing tobacco use EHR content and quality issues. Our results provide insights into the challenges with reconciling and integrating this data for secondary uses. The results indicate the structured tobacco use data alone like do not provide complete tobacco use information. Further work is needed to improve approaches for integration of tobacco use information from different parts of the EHR.

Acknowledgements

The National Institutes of Health through the National Library of Medicine (R01LM011364 and R01GM102282), Clinical and Translational Science Award (8UL1TR000114-02) supported this work.

References

Mental Status Documentation: Information Quality and Data Processes

Charlene Weir, PhD, 1,2 Bryan Gibson, DPT, PhD 2 Teresa Taft, BS1,2 Stacey Slager2 MS, Lacey Lewis1, MS and Nancy Staggers, PhD, RN, FAAN2,3

1VA IDEAS Center of Innovation, SLC, UT; 2Department of Biomedical Informatics University of Utah School of Medicine, SLC UT, 3Summit Health Informatics

Abstract

Delirium is a fluctuating disturbance of cognition and/or consciousness associated with poor outcomes. Caring for patients with delirium requires integration of disparate information across clinicians, settings and time. The goal of this project was to characterize the information processes involved in nurses’ assessment, documentation, decision-making and communication regarding patients’ mental status in the inpatient setting.

VA nurse managers of medical wards (n=18) were systematically selected across the US. A semi-structured telephone interview focused on current assessment, documentation, and communication processes, as well as clinical and administrative decision-making was conducted, audio-recorded and transcribed. A thematic analytic approach was used.

Five themes emerged: 1) Fuzzy Concepts, 2) Grey Data, 3) Process Variability 4) Context is Critical and 5) Goal Conflict. This project describes the vague and variable information processes related to delirium and mental status that undermine effective risk, prevention, identification, communication and mitigation of harm.

Introduction

Delirium is defined as a fluctuating disturbance of cognition (memory, language, orientation) and/or consciousness with reduced ability to focus, sustain, or shift attention. Delirium may impact 14–56% of all hospitalized elderly patients, 2,3 and is associated with poor outcomes, including: increased length of stay, increased likelihood of falls and accidents and discharge to a nursing home . Delirium causes diagnostic dilemmas for physicians and due to the need for increased monitoring, is often a significant burden on nursing staff.4

Several factors make delirium a difficult clinical problem. The first is that delirium often goes undetected: the prevalence of undiagnosed delirium was found to be as high as 32-67% on general medical units, 65% in emergency departments 3 and 86% in nursing homes.5,7 Across 22 studies, Steis and Fick found that nurses’ recognition of delirium symptoms ranged from only 26% to 83%.8 Complicating matters is that delirium may be the result of a variety of causes including: urinary tract infections, post anesthesia effects, pain, and iatrogenic effects from medications.9 A confounding issue in dealing with the problem of delirium is the distributed responsibility for care in inpatient settings. For example, a nurse may note that the patient is confused; this information is likely recorded either in a narrative note or in verbal handoffs or both. Physicians will likely not read the note and unless the nurse informs them verbally, then physicians will likely not be aware of the problem. Since fluctuation in alertness is the hallmark of delirium, multiple observations must be recorded to quantify the problem. Some of the communication norms may be implicit, as individuals may assume that when nothing is mentioned, there is not a problem.

Clinical decision support interventions were shown to increase delirium detection rates. However, despite increased recognition, these interventions were relatively ineffective in improving outcomes in patients with delirium.10-14 In some settings, increased detection of delirium did not result in changes in patient care.11,12,15 In studies showing improved outcomes, improvements appear to be associated with a multidisciplinary approach, including enhanced communication and shared awareness of goals across the healthcare team.6,16-18 Current practices regarding nurses’ data collection and communication processes related to mental status are not well understood. This study helps to fill that gap.

In this paper we describe a qualitative analysis using semi-structure interviews with inpatient nurse managers of acute medicine floors across the Veterans Health Administration (VHA). The purpose of this study was to understand how information about patients' mental status (with a focus on delirium) is recorded, transferred,
communicated and tracked. This formative work was intended to inform the future development of a shared information space for the healthcare team to support communication, decision making, action planning and tools such as decision support technology.

**Conceptual Framework**

Successful detection and treatment of delirium depends on the healthcare team sharing information effectively. Effective communication requires not just sharing information, but sharing one’s beliefs regarding the meaning and implications of the information, expectations regarding the distribution of responsibilities and beliefs about the cause. Based on our belief that shared situational awareness drives effective clinical care of individuals’ with delirium, we based this work on Herbert Clark’s theories of communication. Clark proposes that communication and language are for the general purpose of creating “joint action” or activities with “rational organization around a common goal“19-21 (p. 69) Joint activities involve mechanisms for coordination (e.g. establishing definite reference, negotiating responsibility), for establishing common ground (understanding the intentions, roles, responsibilities and knowledge of each other), for accumulating coordination rules over time (through learning, minutes, and policy) and for the ubiquitous need for constant repair of miscommunication.

**Methods**

**Settings and Study Participants**

The Central Institutional Review Board of the VA, as well as the University of Utah and the Salt Lake City VA local IRBs approved the study. A convenience sample of nurse managers from 18 VA medical-surgical patient care units was selected from across the U.S. Four worked in the VA western area, four in the eastern, three in the southern, and one in a central area. Participants ranged in age from 38-57 years with a mean of 43.5 years. Their years of experience in the VA ranged from 6 months to 29 years with the modal experience at 10-13 years. The majority of nurses had bachelor's degrees (n=6) or master's degrees (n=5). Unit selection maximized variability: medicine, dialysis, rehabilitation, telemetry, transplant, and combined medical/surgical. The average census on units ranged from 6-29 (mean = 18.5) patients.

**Setting**

The VA is supported by VISTA, a mature electronic health record or EHR, (Stage 7 HIMSS Analytics) which includes clinical documentation, closed loop medication administration, orders management, decision support, and data warehousing. The clinician-facing component of VISTA, called CPRS, has standard templates for electronic documentation with the capability to create site-specific and unit-specific electronic forms for clinical documentation.

**Interview Guide Development**

The semi-structured interview guide was developed by the authors (CW, NS and BG) to address 1) current practices in managing and communicating patients' mental status changes, 2) decision-making processes for staffing and resource use, and 3) practices for communication between and within the various clinical roles. The semi-structured interview was piloted and refined during interviews with two nurse managers before the study began.

**Procedures**

The team recruited potential participants through phone contact with the site’s nurse executive and email invitations sent directly to participants. After the consenting process was completed, two researchers (BG, NS) completed telephone interviews using the semi-structured interview guide. The specific interview questions are in the Appendix. The interviews lasted 15-20 minutes at the minimum and sometimes up to an hour for a few and a median of 25”. They were recorded, transcribed, verified for accuracy and de-identified by a professional transcriptionist. Transcripts were loaded into Atlas.tiTM for analysis.

**Data Analysis**

The team used an iterative process of qualitative content analysis.22 After group calibrations, each member of the team independently created “pre-codes” or short text that paraphrased the text. The associated quotations were
discussed and aggregated by assigning more abstract codes in an iterative consensual process of group discussion and re-coding. Final thematic categories were identified by grouping quotations, comparing and contrasting content through discussion and network display analysis and final re-reading of the original text for validation. The authors (CW, BG, and TT) jointly reviewed all 18 of the interviews using Atlas.ti™.

Results

We organized the concepts identified in our analysis into five thematic categories referring to mental status information content and processes: Fuzzy Concepts, Grey Data, Process Variability, Context is Critical, and Goal Conflict. Each is described below with representative quotations.

Theme 1. Fuzzy Concepts: References to mental status are generally “fuzzy” and imprecise

References to acute mental status changes range from informal to moderately structured. A variety of terms are used including: confusion, acute confusion, agitation, altered mental status, delirium, dementia, and, at times, more colloquial terms like “wacky” and “sundowners.” We noted significant reluctance by nurse interviewees to use the word “delirium” unless patients were formally diagnosed as having the condition: One nurse said, “Nurses don’t diagnose.” The most common term nurses mentioned was “confusion.”

Use of structured mental status data is limited to designated fields in EHR templates (e.g. nurse admission note for orientation) and is usually non-specific for delirium (e.g., patient is alert and oriented 4). Delirium is not typically formally assessed in general medical surgical units unless orders such as the Richmond Agitation and Sedation Scale (RASS) and the CAM (Confusion Assessment Method) were rarely used. The methods for assessing mental status and the timing at which the assessments occurred were viewed as being a matter of nurses’ individual clinical judgment.

Some institutions used a structured format for handoffs, such as SBAR (Situation, Background, Assessment Recommendation), as a framework for their handoff forms. However, this framework is not specific enough to ensure consistent documentation and communication of mental status: “The oncoming nurse has their blank SBAR form and they fill it out during their verbal handoff.”

Table 1. Subtheme of Theme 1 “Fuzzy Concepts”

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>The term delirium is rarely used.</td>
<td>“Regarding like behavior, you know, as far as being agitated, calm, cooperative, all of those different kinds of things as well…” “Unfortunately a lot of people use dementia…when it’s not always the case.”</td>
</tr>
<tr>
<td>Assessments are non-specific for delirium</td>
<td>Is delirium formally assessed? &quot;Yes, sort of.&quot; “It’s not a specific dementia or confusion assessment tool.” “We just do the mental status exam. So we look at orientation, whether they’re oriented to person, place, time; and then whether they’re awake, lethargic, unresponsive to, sedated…” “We rarely assess somnolence. Honestly, we rarely have patients that fit that criteria so they are more tuned into…alert and oriented and then relate to what the patient is not oriented to.”</td>
</tr>
<tr>
<td>And then on your unit do you have a particular way that you’re defining either delirium or acute mental status changes?</td>
<td>“Acute mental status change would be one. A lot of times what I hear just between nursing staff is a lot of times they’ll talk about sundowner’s.” “I don’t really see the actual word delirium being said, but we just would use the words, he just got confused all of a sudden.”</td>
</tr>
</tbody>
</table>

Theme 2. Grey Data: Information regarding mental status is invisible or difficult to find
Grey data refers to information that is not available in the EHR, either at the individual, ward level, or system level. As a result, this information can’t be easily accessed for decision-making. Information can be difficult to find, either because it is not integrated with other data that makes it meaningful, because it is in a non-useful format, or because it is buried deep within clinical notes. Nurses’ use of templates for documentation makes it difficult for physicians to locate relevant information because they are lengthy, not created to display information “at a glance” and do not allow indicators for critical information. In addition, key information such as the patient’s risk for developing delirium or trends in mental status signs and symptoms tend to be verbal and informal. Often mental status information is not recorded.

Table 2. Subthemes of Theme 2 “Grey Data”

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual Documentation</td>
<td>“..in the nursing notes, the flow sheet and possibly even in the plan of care if there's a need identified.”</td>
</tr>
<tr>
<td></td>
<td><em>About care plans:</em> &quot;There isn’t…it’s pretty…to be honest, it’s a little disjointed.&quot;</td>
</tr>
<tr>
<td>Hidden data collections</td>
<td>“No, it’s just an Excel spreadsheet that people developed.”</td>
</tr>
<tr>
<td></td>
<td>“Well, we have a…we have a process for, but it’s not a formal…every delirium assessment which is one on every patient I co-sign the note, so anybody that scores a two or above I put in the folder, but I make sure my nurses are following up so I have a list going back over the last year on anybody that scored two or higher.”</td>
</tr>
<tr>
<td>Hand-offs vary and are often</td>
<td>“They have like a spreadsheet that they have information on all of our patients. Let me look at the headings and see if we’ve got that [mental status] as one of our headings.”</td>
</tr>
<tr>
<td>unstructured</td>
<td>“. . .my floor that just designed our own hand-off sheet, charge nurse to charge nurse, and it will say who the high acuity patients are; it will say who the heavy patients are.”</td>
</tr>
<tr>
<td>Persistence of paper</td>
<td>“That’s a paper form, but we…”</td>
</tr>
<tr>
<td></td>
<td>“Oh, yes. Well, if there’s an ‘order’ for neuro check, then . . .we have a paper form to fill out. . .”</td>
</tr>
<tr>
<td>Lack of trended data</td>
<td><em>Q. If you wanted to know how many patients had delirium last month on your unit, how would you find that out?</em>  &quot;That’s a very good question. I don’t know if I could tell you that.”</td>
</tr>
<tr>
<td></td>
<td>“If I were just to go into the CPRS, I would have to go back and look at every patient.”</td>
</tr>
</tbody>
</table>

Theme 3 “Process Variability” in measurement, information exchange and documentation processes.

The methods for assessing mental status and the timing at which the assessments occurred were viewed as being a matter of nurses' individual clinical judgment. Mental status information is also communicated in at least 4 ways: (1) in written form as part of the her formal clinical note templates or narrative, (2) verbally (handoffs, phone calls, rounds, hallway conversations) and (3) in electronic forms outside of the EHR (e.g. spreadsheets), and (4) on paper (e.g. Nurse’s paper “brains” which are thrown away after the shift is over). Guidelines for when, where and how to communicate are not standardized and are largely left up to nursing judgment. Nurses sometimes call physicians and then document the phone call in a note with the title "physician notification" or in an “addendum” to another note.

Timing of nursing documentation is often disjunctive with changes in the delirious patient’s condition. Nursing documentation may occur once in 4 hours, once per shift, or mental status may not be documented regularly at all. The assumption is that if a patient’s mental status changed the information would be communicated verbally. Thus, verbal communication often serves as the mode for sharing mental status information.

Table 3: Subthemes and quotes for Theme 3 “Process Variability”

<table>
<thead>
<tr>
<th>Sub Theme</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communicating delirium to physicians may</td>
<td>“So the patient comes in; it was documented that they were alert and oriented. So the next nurse</td>
</tr>
<tr>
<td>or may not be effective.</td>
<td>takes over the patient and finds that the patient is not oriented to time, place and person and</td>
</tr>
<tr>
<td></td>
<td>is actually a little agitated and doesn’t want to stay... Call the</td>
</tr>
</tbody>
</table>
doctor and the doctor comes in and the patient is fine.”
“They are free to call at any time they have a concern.” “We can go to the doctor any time and tell him that he’s really agitated or anxious or confused…”
“I’ve got a physician group that are currently not receptive as much as I would like for them to be to have a collaborative relationship with the nursing group”

Documentation processes vary, are non-systematic and often unstructured.
“there’s not one unified place for multiple disciplines to document on the plan of care... they can be in multiple places, but they’re all within CPRS”
“There’s no way through CPRS as far as I know at this point in terms...to extract that data.”
“If they have specific things that they have to do regarding that patient, they document it in their individual disciplinary notes”

Assumption that delirium will be discussed
“We’ll talk about discharge, which would talk about their delirium if they were confused, that placement.”
We wouldn’t necessarily bring up a diagnosis of delirium, but we would say maybe the patient is confused or whatever and they require a sitter. That’s about as much as would be discussed

Assessment and documentation depend on nursing judgment: no systematic way to collect data.
We don’t have any nurse-driven, you know, protocols or anything like that to identify someone that’s at risk for delirium, but it’s certainly something that the nursing staff looks out for.”
“It’s up to the nurse to write about it.”

Mental status information is not monitored
“But as far as acute mental status changes, it’s not something that’s routinely reported at the facility level.”
“as far as tracking that, I couldn’t even promise that. I wouldn’t know how to say in the last month how many people had it.”

**Theme 4. Context is Critical:** The meaning of mental status information is derived from context, clinical expectations and baseline comparisons.
The interpretation of mental status information requires fully knowing the patient’s condition. Patients at risk are rarely formally identified. Few participants reported protocols for instituting systematic monitoring or nursing interventions for patients at risk (with the exception of ETOH withdrawal protocols). The concept of identifying patients at risk was even difficult to explain to interviewees. Red flags reported by nurses that required enhanced action included a sudden change in mental status (from a normal baseline), a fall, or a behavioral difficulty. Resistance to “diagnosing” delirium was frequent.

Table 4. Subthemes for Theme 4 "Context is Critical"

<table>
<thead>
<tr>
<th>Subtheme</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delirium is a non-specific sign</td>
<td>“[Delirium] is from medications or alcohol withdrawal or it can be from a variety of things, nothing... .it doesn’t mean that they’re infarcting their brain or anything.”</td>
</tr>
<tr>
<td>Baseline is critical</td>
<td>“Get reports from wherever they came from, if it was a different hospital, just to kind of get a baseline”</td>
</tr>
</tbody>
</table>
| Nurses do look for proximal causes of delirium | “The first thing you’re going to do is you’re going to look and see, are there new medications onboard, which is the number one cause of delirium.”
“We’re on a post-surgical floor, usually any mental status change could indicate a clot coming loose or an additional bleed or
Monitoring and Intervention depends on protocols

“We have an ETOH-er who is coming in to be admitted. He’s almost at the 72-hour mark which is a warning that we should be aware of his mental status. “

We’ll initiate close monitoring so we’ll put a staff member like, okay, the sitters close to the nurse’s station, bed alarms

So it depends on the level of outburst. … All of a sudden they’re on dialysis and they’re not responding, that’s a different assessment and a code will be called”

<table>
<thead>
<tr>
<th>Theme 5: Goal Conflict - Tension exists between patient safety requirements, diagnostic dilemmas, nurse’s staffing levels, accountability and cost. These tensions impact information use.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The response to a finding of delirium is largely increased monitoring (with the resultant increased demand for nursing resources) with the goal of maintaining patient safety. Identifying the cause is often secondary and “confusion” may often be viewed as a permanent state like dementia. When physicians are called, it is often regarding a perceived safety threat or behavioral problem, so patients who are confused but less active are not brought to the physician’s attention. Staffing decisions are often based on verbal descriptions of behavior rather than perceptions of risk. The question addressed is often restricted to whether a higher acuity floor is needed or more nurses. The prevalence of the problem and the importance of keeping patients safe creates a significant burden and most units use sitters and other high-cost options.</td>
</tr>
</tbody>
</table>

Table 5. Subthemes for Theme 5 "Goal Conflict"

<table>
<thead>
<tr>
<th>Staffing is difficult because of mental status and behavioral issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The physicians will write it … but a lot of times you struggle trying to make that happen because you don’t have the staffing.”</td>
</tr>
<tr>
<td>“…don’t have to have a sitter, but if the patient is able to fall out of the bed or something, then I do provide a sitter.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The first thing is patient safety, not finding a cause.</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Nurses can put a sitter in there or the physician can order a sitter, so it’s a multi-disciplinary plan to address that patient for safety sake.”</td>
</tr>
<tr>
<td>“Depending on whether or not he’s a safety risk to himself or not; like I said, we’ll initiate close monitoring so we’ll put a staff member like, okay, the sitters close to the nurse’s station, bed alarms.”</td>
</tr>
</tbody>
</table>

Discussion

In this study we conducted interviews of Nurse Managers of inpatient medical units to explore how information about patients’ mental status is recorded, transferred, communicated and shared. Our focus was on policy and normative processes, as we did not observe practice patterns. We found that information processes for mental status data lack consistency, specificity, and meaning. Our most important findings suggest significant variation across settings. In addition, data collection and recording is often disjunctive with the timeframe to act (e.g. notes are written every shift but patients status may change much more rapidly than this), thus leaving critical information to be communicated verbally. This results in increased cognitive load for nurses because notifying physicians requires deliberation. Finally, much of the text about delirium is embedded in narrative note descriptions and did not support systematic processes for tracking patients with delirium at a unit or institutional level for the purposes of stewardship. These results will be discussed below in terms of communication theory, implications for development of cognitive support for nurses and nursing workload.

Problem with Vague Reference and its effects on Communication and Coordination

Managing acute mental status changes require coordination among the clinicians involved in the patient’s care. Achieving that coordination depends on effective communication. Theorists have noted that effective
communication requires a clear reference to the subject or *Definite Reference*. Failure to have definite reference results in a loss of shared focus among a team and a demand for increased effort to interpret and understand data. Because humans are “cognitive misers,” vague references to mental status may rarely get the attention required to disambiguate. In other words, if nurses and physicians refer to mental status using vague terminology (e.g., wacky) and indefinite specification (e.g., timing of onset, specific descriptors), then people fill in the information gaps with their own automatic and potentially inaccurate expectations.

The clinical context of delirium makes more precise communication particularly important. Physicians and others need to be able to correctly interpret the meaning of a nurses’ notation of abnormal mental status in the patient record. Similarly, nurses need to know about the patient’s risk for delirium, possible attributions of cause, and the degree to which managing delirium should be considered in relation to the goals of care in order to organization their assessments. Other research supports our findings that communication about mental status appears to be imprecise, relying on orientation ratings taken once per shift and using verbal and other informal modes and implicit protocols. Finally, the evidence that physicians rarely read nurses’ notes limits the impact of physicians’ therapeutic decision making due to the inadequate mental status information that nurses currently document.

The illusion of effective communication has been well studied. Individuals overestimate the effectiveness of their communication and have difficulty inferring the goals or the meaning of other’s discourse or writing. This phenomenon was found in medical handoffs as well as typical human interaction. Individuals may be especially prone to misunderstandings when they are exchanging information that is not clearly specified. Errors in communication also occur when people are under a high cognitive load, which is commonplace in clinical care settings. Although only suggestive, our findings regarding the variability and imprecision in documentation and communication processes may be contributing factors to the lack of quality of care for patients with delirium.

**Cognitive Support from EHRs**

Our results support a conclusion that current EHRs fail to provide the level of support for shared situation awareness of the patient’s mental status needed for healthcare teams to appropriately identify and treat delirium. Herbert Clark’s communication theory is very relevant in several areas. Establishing *definite reference* or precision regarding how mental status is referenced is probably the most obvious application. It is simply not easy in the distributed clinical setting to fully understand and grasp vague mental status references measured differently by many individuals and inconsistent rules for communication. Mental status terms often are without reference to context and the diversity in goals are assumed. An example is the prevalence of nurses only measuring orientation, when other metrics or descriptors might be more informative. In addition, nurses may need decision support in order to provide explicit communication to physicians, that would include baseline, additional measures and some information to support a shared understanding. about possible causes if they know them when they communicate without feeling like they are unduly responsible. Technology could really assist this process by pulling text references to mental status as in a “heat map” and include mental notation beyond delirium to better exclude dementia as a default and to express other important attributes, such as depression and anxiety, in the shared display. Time course is very important to diagnosis and so the display should have a time course.

One design feature that may improve communication is shared information displays, which offer an “at a glance” view of data from multiple sources, allowing viewers to develop shared situational awareness. Currently in healthcare the most prevalent form of shared information display is the unit’s whiteboard. Xiao et al. found that whiteboard uses include: team attention management (e.g. drawing attention to salient patient factors), communication, problem-solving/negotiation and staff assignment. Research on electronic whiteboards in operating suites and emergency departments has shown that they can facilitate improvements in both administrative and workflow efficiencies as well as team performance.

**Nursing Workload**

The lack of specificity in language and documentation for mental status data has direct implications for the daily planning and organization of nursing staffing. Confusion, wandering, and behavioral outbreaks present significant risk for the patient and increased workload for nurses. Interviewees in this study often expressed concern over how the patient’s mental status impacted staffing and assignment decisions. Nurses need precise and measurable information about mental status in order to inform decisions at the shift and institution level. The tension between costs and staffing requirements may lead to heavier workloads on nurses raising the issue of safety of patient care.
Limitations

This study has several limitations. First, although data was collected from 18 care facilities across the U.S., they were all within only one health care delivery system (the VA). The distribution of respondents mitigates some concerns, but all respondents were working in one information system and under similar nurse staffing organizational systems. In addition, our conclusions are tentative because we did not observe the flow of information directly. Our conclusions are based on results from Nurse Manager interviews; as a result, there is some degree of social desirability bias as well as simple inaccuracies.

Conclusions

Documentation and communication processes for mental status information appear to be variable, informal and imprecise. Interpretation of mental status data for nurses’ clinical decision-making requires information on context standardization and precision in references to altered mental status, and integration of information with goals of care. The ability to provide stewardship decisions may require better data capture, decision support tools, and also better communication processes.

REFERENCES


1226


**Acknowledgments**

**Author Contributions.** All authors listed have contributed significantly to the authorship of this paper. Dr. Weir was the PI of the original funding project and led the project. Dr. Gibson participated in all aspects of the project, including interviews and qualitative analysis. Teresa Taft contributed in the qualitative analysis and the write-up of the paper. We appreciate the work of Robyn Barrus in helping with the qualitative software.

**Conflict of Interest.** None of the authors listed have any conflict of interest.

**Sponsor Role.** The funding agency, VA HSR&D allowed full independence in the conduct of this research. The following table lists individual areas of conflict.

**Funding acknowledgement:** Veterans Health Administration Health Services Research & Development: # CRE 12-321
A Mixed Methods Task Analysis of the Implementation and Validation of EHR-Based Clinical Quality Measures

Nicole G. Weiskopf, PhD,1 Faiza J. Khan, MBBS, MBI,1,2 Deborah Woodcock, MBA,1 David A. Dorr, MD, MS,1 Joaquin E. Cigarroa, MD,2 Aaron M. Cohen, MD, MS1
1Department of Medical Informatics & Clinical Epidemiology, OHSU, Portland, OR
2Knight Cardiovascular Institute, OHSU, Portland, OR

Abstract
Clinical quality measures (CQMs) are important tools for the assessment and improvement of health care quality. Federal requirements initially set forth in the American Recovery and Reinvestment Act, and advanced in subsequent stages of the requirements, codified electronic health record (EHR)-based CQM reporting, and have made automated CQM implementation a priority amongst the clinical and informatics communities. Nevertheless, the processes surrounding CQM implementation and validation remain complex, time-consuming, and largely undefined. We collected issue-tracking data during the course of an agile and rigorous collaborative project to build an analytics platform for the Knight Cardiovascular Institute at OHSU, with nine heart failure CQMs defined by the American College of Cardiology (ACC) as an exemplar. Using a mixed methods approach we provide an overview of our CQM implementation and validation process, identify major roadblocks and bottlenecks, and make recommendations for other professionals working in the area of health care quality assessment and improvement.

Introduction
Health care quality is commonly understood to derive from a combination of structures, processes, and outcomes in the clinical setting.1 Measurement of these factors is enabled through the use of clinical quality measures (CQMs), which are used to calculate the proportion of relevant cases (usually patients or visits) that meet the target criteria. CQMs are an essential component of the quality improvement process. The passage of the American Recovery and Reinvestment Act in 2009 introduced the Health Information Technology for Economic and Clinical Health Act, which incentivized the adoption of electronic health record (EHR) systems meeting meaningful use requirements, including the ability to submit CQMs.2

CQM implementation exists on a continuum of automation. Some CQMs are only available in human-readable format, while others, called electronic CQMs (eCQMs), which are intended to be implemented electronically, are available in computer-readable format and are explicitly designed to interface with EHR data and systems. Human readable CQMs, including the measures discussed in this paper, require manual interpretation and mapping to the appropriate clinical concepts. They are not designed specifically for electronic data sources or information systems. It is possible, however, to manually translate human-readable CQM logic into computerized queries and map the clinical concepts to electronic data fields. This electronic implementation of human readable CQMs offers a number of advantages over manual calculation. After initial implementation, for example, calculation of EHR-based automated CQMs is less time- and resource-intensive. Also, automated CQMs can be calculated on an ongoing basis, allowing for ongoing monitoring of care provision and the detection of temporal trends.

Unfortunately, the electronic implementation of human-readable CQMs is non-trivial. The process of interpreting a CQM specification, mapping it to a data source, and demonstrating validity of the measure can be difficult and time-consuming. Few resources exist to guide this process. Moreover, it has been shown that the quality of clinical data underlying CQMs and eCQMs is often poor,3 making it difficult to differentiate between implementation problems and data quality problems at the validation stage. Studies of the implementation of computer-readable eCQMs, which should be more straightforward than electronic implementation of human-readable CQMs, have revealed significant difficulties and roadblocks. A study commissioned by the American Hospital Association (AHA), identified a number of challenges in eCQM implementation and calculation, included difficulty in using eCQM specifications, problems with existing eCQM tools, and issues surrounding the documentation of relevant clinical information in a format accessible by the eCQMs.4 Another study, which looked at five eCQMs from the National Quality Forum (NQF), found problems with the completeness, accuracy, and overall computability of the measures. The authors determined that implementing eCQMs required substantial expertise and manual development.5 It can be inferred that the implementation of human-readable CQMs is even more fraught with challenges.
Over the past year and a half we have used an agile, collaborative approach to implement a set of nine human-readable cardiovascular CQMs as electronic, EHR-based CQMs. By analyzing data from a customized issue tracking database, our intent is to uncover actionable information regarding the processes and workflows involved in this project. What categories of work took more or less time than expected? What were the bottlenecks? What recommendations would we make for planning a similar project going forward?

Methods

Setting

Since July 2014, the Informatics Discovery Laboratory (IDL), an agile, results-driven program within the Department of Medical Informatics and Clinical Epidemiology at Oregon Health & Science University (OHSU) has partnered with the Knight Cardiovascular Institute (KCVI) at OHSU in order to develop, evaluate, and maintain an analytics platform capturing information related to quality of care and health care delivery. The multidisciplinary IDL team included informaticians, a data architect, a developer, a project manager, a data analyst, and research assistants, dedicating to the project a combined total of 2.35 full-time equivalents in the first fiscal year, and 2.81 full-time equivalents in the second fiscal year. The pilot project of this collaboration, which was completed in February 2016, focused on the implementation of nine heart failure CQMs, according to the specifications of the American College of Cardiology (ACC). The nine measures are summarized in Table 1.

Table 1. Summary of ACC heart failure CQMs. *LVEF: Left Ventricular Ejection Fraction; **ACE: Angiotensin-converting-enzyme inhibitor; ARB: Angiotensin Receptor Blocker; +ICD: Implantable Cardioverter Defibrillator

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LVEF* Assessment (outpatient)</td>
</tr>
<tr>
<td>2</td>
<td>LVEF Assessment (inpatient)</td>
</tr>
<tr>
<td>3</td>
<td>Symptom Assessment</td>
</tr>
<tr>
<td>4</td>
<td>Symptom Management</td>
</tr>
<tr>
<td>5</td>
<td>Patient Education</td>
</tr>
<tr>
<td>6</td>
<td>Beta-blocker Therapy</td>
</tr>
<tr>
<td>7</td>
<td>ACE/ARB** Therapy</td>
</tr>
<tr>
<td>8</td>
<td>ICD* Counseling</td>
</tr>
<tr>
<td>9</td>
<td>Post-Discharge Appointment</td>
</tr>
</tbody>
</table>

The analytics platform was built upon an existing health information technology platform, the Integrated Care Coordination Information System (ICCIS), which includes a robust CQM calculation and reporting engine. ICCIS was developed to support Care Management Plus, a care coordination, quality improvement, and information technology model developed for use by primary care clinics.

Design and Data Collection and Management

Early in the process of implementing and evaluating the ACC measures, the IDL recognized the need for an issue tracking and change management system, leading to the development of a specialized Microsoft SharePoint site. This site allowed collaborators on the project to open, modify, assign, update, and close tasks, decisions, requests, and issues. Each task could be categorized as belonging to a specific measure, or belonging to all the heart failure measures. Following the completion of the heart failure CQM project, we extracted the following fields from the SharePoint site: task name, subproject (measure), date opened, last date modified, and status (open, closed, or noted). Once a task was opened it was not closed until it was completed, even if the work was temporarily put on hold. It must therefore be emphasized that task durations were more representative of how long it took to complete a task than how much time was actually spent working on a task. It should also be noted that many of the tasks overlapped significantly, and had complex interdependencies.

The dataset underwent a number of cleaning steps. First, we eliminated decisions and future issues, keeping only those item types that required time during the study period to complete. Second, we limited the dataset to
completed items. Third, we combined ACC measures 6 and 7 (beta-blocker and ACE/ARB therapy, respectively), since the two measures are intended to be paired and much of the work on them was performed in parallel. Fourth, due to missing tasks and completion dates related to project management, we removed all project management issues from the dataset. Finally, some work had to be done to identify the completion dates of several items that had unintentionally been left open on the SharePoint site. Missing completion dates were located by searching project emails and metadata from files and ICCIS.

Analysis

We used a general inductive approach to derive and apply category labels based on the issue tracking data. Three of the investigators used an iterative, open coding approach to create a set of exhaustive and mutually exclusive categories to capture the different types of work involved in the development and evaluation of the measures. The final category for each task was determined through consensus. For each task we calculated the number of days from date opened to final date modified to estimate the amount of time required for completion. We referred to this metric as task-days, in order to emphasize the possibility of concurrent rather than purely sequential work. Overall task-days were summed across measures and across task categories. The duration of the work for each measure was also calculated. Basic summary statistics were calculated for each category both within and across each of the measures. Gantt charts were used to visualize the temporal relationships between tasks.

Table 2. We derived eight categories of work necessary for implementing, validating, and utilizing EHR-based CQMs. The categories are listed from top to bottom in roughly expected order of occurrence.

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition and Examples</th>
<th>Task count</th>
<th>Task-days</th>
<th>Mean task-days (SD)</th>
<th>Percent of task-days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation</td>
<td>Interpretation and operationalization of CQM concepts, target population, and related issues:</td>
<td>7</td>
<td>301</td>
<td>43.0 (38.2)</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>• Generate list of evidence-based beta-blockers and ACE/ARBs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Define a KCVI inpatient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data exploration</td>
<td>Identification and selection of appropriate data fields and values:</td>
<td>13</td>
<td>842</td>
<td>64.8 (60.9)</td>
<td>13.5%</td>
</tr>
<tr>
<td></td>
<td>• Find follow-up appointment data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Find medication exception reasons</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System development and debugging</td>
<td>Development, maintenance, correction, and updating of back-end data capture and pre-processing system:</td>
<td>45</td>
<td>1868</td>
<td>41.5 (45.6)</td>
<td>29.9%</td>
</tr>
<tr>
<td></td>
<td>• ETL for smart data elements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Add missing providers to provider table</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measure development and debugging</td>
<td>Development, maintenance, correction, and updating of measure queries and measure-specific programming:</td>
<td>19</td>
<td>1463</td>
<td>77.0 (81.6)</td>
<td>23.4%</td>
</tr>
<tr>
<td></td>
<td>• Write numerator query for measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Debug queries for measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>Determining quality of automated CQMs and true performance based on manual chart review:</td>
<td>14</td>
<td>594</td>
<td>42.4 (27.3)</td>
<td>9.5%</td>
</tr>
<tr>
<td></td>
<td>• Create validation spreadsheet for measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Perform validation for measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synthesis and analysis</td>
<td>Quantitative, qualitative, and graphical analysis of automated CQMs:</td>
<td>12</td>
<td>1048</td>
<td>87.3 (78.5)</td>
<td>16.8%</td>
</tr>
<tr>
<td></td>
<td>• Update visualizations with new measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Conduct error analysis of validated measure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informing and updating stakeholders</td>
<td>Delivery of findings and recommendations to stakeholders:</td>
<td>6</td>
<td>124</td>
<td>20.7 (22.1)</td>
<td>2.0%</td>
</tr>
<tr>
<td></td>
<td>• Get feedback on visualizations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Present quarterly review</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>116</strong></td>
<td><strong>6240</strong></td>
<td><strong>54.3 (59.3)</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
Results

We derived seven categories of tasks from the issue tracking database, which included a total of 116 tasks in the final dataset, as summarized in Table 2. The total number of tasks per category ranged from six (informing and updating stakeholders) to 45 (system development and debugging). The time required per task varied greatly; the mean was 54.5 days, with a standard deviation of 112.9 days. The shortest tasks, on average, were related to informing and updating stakeholders, and the longest were related to synthesis and analysis. The issue tracking database also included 15 tasks related to project management, which were not recorded at the measure-level.

The complete, end-to-end implementation, validation, and reporting of each of the ACC measures required, on average, 459.9 task-days, lasted an average of 265.8 calendar days, and had an average of 8.4 tasks. The overall distribution of tasks and durations by category and measure are shown in Figure 1. The number of recorded tasks for each measure ranged from three (the symptom assessment, symptom management, and patient education measures) to 21 (the paired medication measures). The work on the paired medication also took the greatest number of task-days total to complete (955). The patient education measure (measure 5) took the fewest task-days (54). Complete measure-level summary statistics are presented in Table 3.

Table 3. A summary of the work conducted for each measure. “All measures” in this case denotes tasks that were assigned to all measures, not a combination of tasks related to individual measures.

<table>
<thead>
<tr>
<th>ACC HF Measure</th>
<th>Task count</th>
<th>Duration</th>
<th>Task-days</th>
<th>Percent of task-days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LVEF* Assessment (outpatient)</td>
<td>11</td>
<td>329</td>
<td>295</td>
<td>4.7%</td>
</tr>
<tr>
<td>2 LVEF Assessment (inpatient)</td>
<td>11</td>
<td>333</td>
<td>645</td>
<td>10.3%</td>
</tr>
<tr>
<td>3 Symptom Assessment</td>
<td>3</td>
<td>222</td>
<td>250</td>
<td>4.0%</td>
</tr>
<tr>
<td>4 Symptom Management</td>
<td>3</td>
<td>222</td>
<td>250</td>
<td>4.0%</td>
</tr>
<tr>
<td>5 Patient Education</td>
<td>3</td>
<td>28</td>
<td>54</td>
<td>0.9%</td>
</tr>
<tr>
<td>6&amp;7 Beta-blocker and ACE/ARB Therapy</td>
<td>21</td>
<td>365</td>
<td>955</td>
<td>15.3%</td>
</tr>
<tr>
<td>8 ICD+ Counseling</td>
<td>6</td>
<td>300</td>
<td>343</td>
<td>5.5%</td>
</tr>
<tr>
<td>9 Post-Discharge Appointment</td>
<td>9</td>
<td>327</td>
<td>887</td>
<td>14.2%</td>
</tr>
<tr>
<td>All measures</td>
<td>48</td>
<td>434</td>
<td>2561</td>
<td>41.0%</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>447</td>
<td>6240</td>
<td>100.0%</td>
</tr>
<tr>
<td>Mean</td>
<td>8.4</td>
<td>265.8</td>
<td>459.9</td>
<td>7.4%</td>
</tr>
</tbody>
</table>

Figure 2 shows the overall workflow of the project, broken down by task category and measure. This highlights the nonlinear nature of the approach, with several measures requiring the development team to step back to a previous area. For instance, measure 1, outpatient LVEF assessment, required several repeated efforts in system development and debugging in order to identify the correct patients and the correct visits within the ICCIS system. Measure 9, post-discharge appointment scheduling, required extensive data exploration to locate and extract follow-up appointment information, which caused delays in measure development and debugging.

Discussion

We completed the implementation of nine clinical quality measures related to cardiovascular care from standard EHR data; this task took, on average, 459.9 task-days of effort and 8.4 separate tasks per measure. We highlight the work categories, the iterative nature of the process, and challenges below.

Derived work categories and process

Our work focused on the electronic implementation of traditional (human readable) CQMs, rather than computable eCQMs, but it is still informative to compare our process to published work on eCQM implementation. The AHA, for example, proposes an iterative eCQM implementation process involving five major steps. That process includes: “gap analysis,” which can be mapped to Data exploration; “data extraction and eCQM calculation,” which is related to System and Measure development and debugging; “validation;” and “downstream uses of eCQM results,” which is largely undefined but could be said to include synthesis and analysis and informing and updating stakeholders. The AHA framework also includes “data capture and workflow redesign” within the implementation and validation iteration, which we see as a future step in our process. Therefore, despite the fact that eCQMs are meant to be more straightforward to implement than non-computable CQMs like the ACC measures, the only category of work we captured in the issue-tracking data that is not included in the AHA framework is the actual...
interpretation of each measure. Similar frameworks, such as that from the Office of the National Coordinator (ONC) for HIT, also have the organizational structure required highlighted, which was particularly important in our process.

**Figure 1.** Number of tasks and time spent on tasks in each category, by measure. Tasks assigned to all measures are not included here.

The most significant difference between our derived process and those described by the AHA and ONC is the inclusion of a step focused on workflow and documentation changes and improvements, which was outside the scope of the work we have completed to date. Our work, however, has informed an ongoing project focused on improving documentation in order to improve adherence to clinical guidelines and the accuracy of the ACC heart failure CQMs. As informaticians, we found we could accomplish a great deal prior to this necessary set of steps; now, other collaborators can move forward with workflow and documentation changes after current data quality and extraction have been optimized.

**Nonlinearity of implementation process**

Initially, we had anticipated that the CQM implementation and validation process would be largely sequential, with certain areas of iteration around system development and debugging and measure development and debugging. Because we were building on the existing, robust ICCIS system, we expected that there would be minimal system development and debugging, and that what we did find would be concentrated near the beginning of the project. We expected that each measure would follow a reasonably predictable progression, starting with interpretation of the measure specifications, and ending with efforts to inform and update stakeholders. The inpatient LVEF measure (measure 2), as visualized in Figure 2, roughly follows this linear process. Some of the other measures, however, including outpatient LVEF assessment, ICD implantation counseling, and follow-up appointment scheduling (measures 1, 8, and 9, respectively) followed a less structured, more iterative path, requiring extended work related to system development and debugging and data exploration. Tasks that applied to all measures were similarly nonlinear. The difference between the expected process and the observed process is visualized conceptually in Figure 3. At any stage of the process it was possible to discover an issue that required returning to any of the preceding stages. Measure debugging, validation, or synthesis and analysis, for example, might uncover system errors, like problems with data flow or population selection. System development and debugging, in turn, might reveal that a data element was being used differently than originally anticipated, which would require further data exploration to find a better element, or even interpretation, to select a similar concept that might be documented more consistently. In general there were frequent returns to system development and debugging, in order to customize the ICCIS system to the requirements of the KCVI project.

**Example roadblocks and challenges**

EHR data are often fragmented and unstructured, which makes it difficult to map concepts from CQMs to elements within the EHR, as well as incomplete or otherwise unreliable, which necessitates more complex
<table>
<thead>
<tr>
<th>ACC 1</th>
<th>Interpretation</th>
<th>Data exploration</th>
<th>System development and debugging</th>
<th>Measure development and debugging</th>
<th>Validation</th>
<th>Synthesis and analysis</th>
<th>Informing and updating stakeholders</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC 2</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
<tr>
<td>ACC 3</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
<tr>
<td>ACC 4</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
<tr>
<td>ACC 5</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
<tr>
<td>ACC 6 &amp; 7</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
<tr>
<td>ACC 8</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
<tr>
<td>ACC 9</td>
<td>Interpretation</td>
<td>Data exploration</td>
<td>System development and debugging</td>
<td>Measure development and debugging</td>
<td>Validation</td>
<td>Synthesis and analysis</td>
<td>Informing and updating stakeholders</td>
</tr>
</tbody>
</table>

Figure 2. By-measure workflows. Concurrent tasks within categories will overlap.
Figure 3. Expected process, which is mostly linear with specific instances of iteration at the development and debugging stages, and observed process, which may include substantial iteration.

measure logic. The ACC heart failure measure specifications refer to CPT Category II codes that capture specific clinical problems and actions necessary for the computation of these measures. Our institution, however, does not utilize these codes, which means that we instead had to identify, extract, and map other sources of this information. As shown in Figure 2, the paired medication measures (measures 6 and 7), which quantify clinician performance on the prescription of evidence-based beta-blockers and ACE/ARB for patients with a history of decreased systolic function, required substantial iterative work related to measure development and debugging and validation, largely because of EHR documentation limitations. After beginning to implement the measures based on medication data, we learned that clinicians documented medication adherence and medication exceptions in two other structured fields, depending upon setting. These fields, however, were not always used consistently, were unreliable, and were sometimes out of date; some patients who were labeled as having exceptions were in fact on the appropriate medications. To account for this data quality problem we needed to develop and evaluate significantly more complex logic.

Another unexpectedly complex problem involved the identification of relevant patients and the attribution of those patients to the appropriate providers for group- and provider-level performance calculation and reporting. Many of the system development and debugging tasks captured in the outpatient and inpatient LVEF measures (measures 1 and 2) and under all measures related to this work. In implementing the LVEF measures we initially had to develop system rules and processes for the identification and extraction of relevant patients and their data, which were different for the outpatient and inpatient populations. In a series of tasks related to all measures we also addressed how these patients were, in turn, “assigned” to providers for performance calculation. There are numerous models of attribution in CQM systems.15, 16 Two common approaches are to attribute patients to primary care providers or the most recent care provider. These approaches, however, may not be appropriate in settings where care is delivered by teams or in settings where providers frequently see referral patients. Instead, we used a multiple attribution rule, where all providers who saw a patient during the measurement period received “credit.” Implementing this rule, even within a robust CQM system like ICCIS, involved substantial technical effort.

A substantial amount of synthesis and analysis was performed through the course of iterative efforts to develop visualizations and reports that provided the KCVI stakeholders with the type and granularity of information they needed, as captured in the Figure 2 workflow for all measures. Many different versions of the reports were
developed, tested, and reworked before arriving at a final product that was information-dense, comprehensible, and actionable.

Overall, in the course of implementing and validating the nine ACC heart failure measures we identified a number of significant roadblocks, often involving complexities of EHR documentation and patient identification and performance attribution. We also discovered that CQM implementation is an unavoidably iterative process, with seven distinct stages of work.

Limitations

The issues-tracking data described in this paper were a byproduct of an operational, real-world, agile process, and were therefore not collected with the same care as traditional research data. Therefore, there was a certain degree of inconsistency in how and when the issues were opened and closed. Moreover, many tasks were left open for an extended period while no actual work was being done, usually because of task interdependencies or because other tasks were taking priority. Both of these factors could impact the calculated task durations. Other tasks, especially those that were straightforward or could be completed rapidly, were not tracked at all. Many of the simpler measures, therefore, had incomplete tracking data. We expect, therefore, that our final dataset suffered from some degree of inaccuracy and incompleteness.

One specific problem resulting from incomplete tracking data was limited tracking of tasks related to project management. The CQM implementation efforts required significant administrative and organizational efforts that were not reflected in our data. Examples of such tasks include training, recruitment, infrastructure and tool development, communication, and project planning. The work done in this area was significant and vital for the execution of the overall project, and future research should be done to fully understand the role of project management in successful analytics development and implementation.

It is also important to note limitations in generalizability of our findings due to setting and clinical focus. Data elements relating to heart failure, including relevant assessments and patient care steps, are unique and may therefore present challenges that would not be seen in all clinical areas or subspecialties. We also faced unique challenges in implementing these measures for a cardiovascular institute providing both inpatient and outpatient services within an academic tertiary care medical center. Nevertheless, we believe that our findings and be extendable and informative for other CQM implementers in diverse settings.

Conclusion

Through a mixed methods analysis of issue-tracking data, we derived a set of seven categories of work relating to the end-to-end implementation and evaluation of valid and reliable automated EHR-based CQMs. These align well and expand on prior work in this area. We encountered a number of expected and unexpected challenges during this work, which were captured by these data. These challenges derived largely from EHR documentation limitations, and from back-end and calculation-related challenges, and new features needed to support data analytics. We also encountered substantial iteration, which was, in some cases, not only unavoidable, but also a vital feature of a complex development and knowledge discovery process. In other cases, this iteration could have been avoided through an exhaustive exploration of user needs and documentation practices at the start of the CQM implementation project.

Funding

The work was supported by the Knight Cardiovascular Institute at Oregon Health & Science University.

References

Differentiating Sense through Semantic Interaction Data

T. Elizabeth Workman, PhD, MLIS1, 2; Charlene Weir, PhD, RN1, 3; Thomas C. Rindflesch, PhD4

1VA Salt Lake City Health Care, Salt Lake City, Utah
2Division of Epidemiology, University of Utah, Salt Lake City, UT
3Department of Biomedical Informatics, University of Utah, Salt Lake City, UT
4Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Words which have different representations but are semantically related, such as dementia and delirium, can pose difficult issues in understanding text. We explore the use of interaction frequency data between semantic elements as a means to differentiate concept pairs, using semantic predications extracted from the biomedical literature. We applied datasets of features drawn from semantic predications for semantically related pairs to two Expectation Maximization clustering processes (without, and with concept labels), then used all data to train and evaluate several concept classifying algorithms. For the unlabeled datasets, 80% displayed expected cluster count and similar or matching proportions; all labeled data exhibited similar or matching proportions when restricting cluster count to unique labels. The highest performing classifier achieved 89% accuracy, with F1 scores for individual concept classification ranging from 0.69 to 1. We conclude with a discussion on how these findings may be applied to natural language processing of clinical text.

Introduction

Word Sense Disambiguation (WSD), a common task in Natural Language Processing (NLP), is the process of determining the precise meaning of an ambiguous word for a given instance in text. For example, the word “cold” in biomedical text can refer either to rhinovirus (e.g., “the common cold affects many people”) or cold temperature (e.g., “the patient felt cold in the room”). In this task, we have a single representation “cold” that can possess one of multiple, distinct definitions, depending on the use of the word. WSD is usually achieved through analyzing the context in which the word appears1. Recent popular biomedical WSD applications include disambiguation of terms2, 3 plus abbreviations and acronyms4.

Another difficult but similar task to WSD is to differentiate two terms that have different representations, but similar meanings. For example, dementia and delirium are similar in manifestation, but they are separate, independent conditions. Delirium or Acute Changes in Mental Status is defined as a fluctuating disturbance of cognition (memory, language, orientation) and/or consciousness with reduced ability to focus, sustain, or shift attention5. Delirium may impact 14–56% of all hospitalized elderly patients6, 7, and is associated with poor outcomes, including: increased length of stay, increased likelihood of falls and accidents and discharge to a nursing home. Delirium causes diagnostic dilemmas for physicians often trying to discriminate delirium from dementia in elderly patients, and due to the need for increased monitoring, is often a significant burden on nursing staff8. They are different clinical events and this poses important implications in treatment and other aspects of case management. However, there are inconsistencies in even provider representations. Researchers found that clinicians inconsistently document delirium in veteran electronic health records, even when there is a confirmed diagnosis9. Additionally, they use alternative terms such as “disoriented”, “muttering”, and “showing confusion” in documenting delirium10. Inconsistent documentation in general causes clinical care team members to expend extra effort to validate data11. Prolonged delirium in care settings poses significant risks of cognitive impairment and death12. Therefore, it is vital to identify and treat it quickly. However it is often not recognized, or misdiagnosed as dementia or another psychiatric condition13. Dementia recognition is likewise important, yet early-onset dementia is often misdiagnosed, which can have devastating consequences on patients and their families14.

Data that capture how semantic elements interact in text may help explicate the differences between two separate terms that have similar meanings. This may be especially useful in distinguishing incidents of dementia or delirium in text. This study explores the application of such data to clustering and classifying exercises, as an initial effort to explore the use of it to differentiate two related terms and other NLP tasks.
Background

Semantic predications provide a representation of biomedical text that has been distilled to simple assertions, consisting of the semantic elements expressed in the original text. SemRep\textsuperscript{15}, an NLP application, extracts semantic predications from PubMed title and abstract text. For example, SemRep takes the following text:

“…calcium channel blockers…reduce the risk of dementia…”\textsuperscript{16}

and extracts this semantic predication:

Calcium Channel Blockers (Pharmaceutical Substance) PREVENTS Dementia (Disease or Syndrome)

SemRep identifies the two prominent arguments in the text and their preferred terms from the UMLS Metathesaurus (C0006684 Calcium Channel Blockers and C0497327 Dementia), and the predicate, or relationship that binds them (PREVENTS) as indicated in the UMLS Semantic Network. SemRep also identifies the semantic types, or subject classes of the two arguments, indicated in parentheses. To better define this semantic predication topologically, one could refer to Calcium Channel Blockers as the opposite argument to Dementia, and visa-versa.

There is a resource that provides semantic predications for research and other purposes. The Semantic MEDLINE database\textsuperscript{17} contains over 84 million semantic predications extracted by SemRep from the biomedical literature from 1900 to 2015.

These data structures may shed light for differentiating delirium and dementia; however, a basic analysis demonstrates substantial contextual similarities between the two terms. Semantic predications containing either term share many co-occurring elements. Delirium, the smaller in terms of total unique semantic elements with which it occurs, shares all of its predicates, 54% of its opposite arguments, and 98% of its opposite argument’s semantic types with Dementia, in semantic predications containing both terms extracted from literature published from 2005 - 2014.

An analytical method such as clustering may provide differentiation between delirium and dementia as they appear in semantic predications. Expectation Maximum (EM) clustering\textsuperscript{18} was designed to identify latent, or implicit variables from mixture models of shared probability density functions. Concept data drawn from the Semantic MEDLINE database can be characterized as representing a mixture model, where each concept belongs to subpopulations in which it interacts with other semantic elements. To activate instances of the EM algorithm, one must prescribe an amount of clusters presumed to be representative of the data. Algorithms such as Bayesian Information Criterion (BIC) determine an optimal cluster count which then enables the EM clustering process.

Frequency data and its membership in mixture models has previously been exploited in classification\textsuperscript{19}. EM Clustering has previously been applied in a semantic clustering context\textsuperscript{20}, including efforts for building sense inventories for abbreviations in clinical notes\textsuperscript{21}. This work applies EM Clustering and other machine learning techniques to data extracted from the Semantic Medline database.

Objectives

The purpose of this work is to determine if (a) interactive rate data follow a clustering behavior expectant of data drawn from a mixture model, and (b) this data could successfully be used to classify concept sense. Applying interaction values, i.e., representations of how each concept interacts with other semantic elements, to a clustering method, might potentially identify clusters of expected proportions. These values might also be used to train an effective classifier to identify instances of several concepts. Specifically, we wanted to answer the questions:

- Do unlabeled datasets of semantically related concept pairs exhibit two distinct clusters of equal or similar proportions when subjected to BIC modeling and EM clustering?
- When labels are included, how do data instances cluster?
- Could a dataset for several concepts, where there is matching data for another semantically related concept for each concept in the dataset, be applied to training several different types of classifiers? If so, what classifier would perform best?

To explore these questions, we analyzed data from several semantically related concept pairs including delirium and dementia. We first subjected datasets of concept pairs to two clustering processes, the first without the concept labels, then the second with the labels restored. We then applied the data for all concepts to training and testing several classifier algorithms.
Methods

Data Procurrence and Preparation

We gathered the following features for each term for a ten-year period (2005-2014) from the Semantic MEDLINE database: frequency of the term appearing as an argument in the database, count of unique terms appearing in the opposite argument position, count of unique semantic types of the terms in the opposite argument position, count of unique predicates, count of unique semantic type-predicate combinations, and count of unique opposite argument-predicate combinations. We retrieved this data according to date of entry (EDAT) into the PubMed database. We aggregated it in one-month intervals to form vectors for these six features in order to provide many instances for the clustering and classifier training tasks. The following example is an instance of this data for Dementia. The features are concept label, concept occurrence frequency, unique opposite argument frequency, unique predicate frequency, unique semantic type frequency of the opposite argument, unique semantic type-predicate combinations, and unique opposite argument-predicate combinations:

Dementia, 160, 39, 12, 19, 29, 45

In other words, for this given month, there were 160 instances of the concept “Dementia”, 39 unique concepts appearing with it in the opposite argument position, 12 unique predicates, 19 unique semantic types represented among the opposite argument concepts, 29 unique semantic type-predicate combinations, and 45 unique opposite argument-predicate combinations. We gathered this data for the following concept pairs (Table 1):

<table>
<thead>
<tr>
<th>Concept 1 Name</th>
<th>Concept 1 CUI</th>
<th>Concept 2 Name</th>
<th>Concept 2 CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delirium</td>
<td>C0011206</td>
<td>Dementia</td>
<td>C0497327</td>
</tr>
<tr>
<td>Heart</td>
<td>C0018787</td>
<td>Myocardium</td>
<td>C0007061</td>
</tr>
<tr>
<td>Abortion</td>
<td>C0156543</td>
<td>Miscarriage</td>
<td>C0000786</td>
</tr>
<tr>
<td>Congestive Heart Failure</td>
<td>C0018802</td>
<td>Pulmonary edema</td>
<td>C0034063</td>
</tr>
<tr>
<td>Stroke</td>
<td>C0038454</td>
<td>Infarct</td>
<td>C0021308</td>
</tr>
<tr>
<td>Delusion</td>
<td>C0011253</td>
<td>Schizophrenia</td>
<td>C0036341</td>
</tr>
</tbody>
</table>

Table 1. Concept Pairs

The additional pairs were rated as significantly semantically related by three clinicians in a former study. All but one concept yielded 120 instances. Abortion (C0156543) yielded 116 instances, or, in other words, there were insufficient data for four months of the 10-year collection period to produce data for this study. We did not use any imputation methods to create data for the four missing instances in this dataset.

After retrieving the data, we normalized the values of each quantitative feature for each concept by scaling on a mean of 0 and a standard deviation of 1. For each concept dataset, each quantitative feature was represented in a single vector, forming a total of six vectors with 120 values each, for each dataset, except for the Abortion dataset, where there were six vectors with 116 values each. The corresponding concept label formed an additional feature vector in each dataset; these labels were excluded for the first clustering exercise (described in the following section). For classifier training, we also randomized the instances and split the dataset into a training subset (80%) and a testing subset (20%).

Clustering Analysis

We applied a model-based clustering approach suggested by Fraley using the MCLUST software package in R. This technique implemented BIC in combination with EM Clustering and hierarchal agglomeration to find and execute the best model (including cluster count and characteristics) for a given dataset. This is done under the premise that the data in question is drawn from a mixture of multiple probability distributions. In addition to cluster count, output also includes cluster characteristics such as volume, shape, and orientation, according to the best model. We also documented the covariance matrix model implemented in determining the best BIC model (and cluster count). We analyzed each dataset for each concept pair, removing the concept labels from the instances before processing.

The software also included functionality to cluster labeled data for cluster classification by applying Eigenvalue Decomposition Discriminant Analysis. We applied this process to the concept pair data while including concept class labels. This exercise allowed us to view how the concept pair data instances divided into categories when matched against possible classifications.
Classifier Training

Using the data for all 12 concepts, we trained several types of classifiers, specifically a Support Vector Machine, a Random Forest (i.e., tree structure), a Neural Network, and a Naïve Bayes classifier, because these are identified as commonly used in the related task of WSD\textsuperscript{26}. We employed the following R packages for the four classifier types, to create basic versions of their resulting products:

- Support Vector Machine: libSVM (in the e1071 package)\textsuperscript{27}
- Random Forest: randomForest\textsuperscript{28}
- Neural Network: nnet\textsuperscript{29}
- Naïve Bayes: e1071\textsuperscript{27}

This was a simple, exploratory exercise, to analyze the results from a variety of classifiers. In addition to the training data, we added the following parameters to each classifier instance:

- SVM: A cost value of 100; Gamma value of 1.
- Random Forest: Importance of predictors assessed; 2000 trees produced.
- Neural Network: Five neurons in a single hidden layer; initial weight value of 0.1; weight decay of 0.00005; 200 maximum iterations.
- Naïve Bayes: (all application default settings implemented)

Otherwise, the basic default parameters provided by each application were used.

Results

Clustering Analysis

Of the six concept pairs, all but one exhibited two clusters of similar or equal proportions, in terms of data elements (Table 2) when clustering data without their concept labels. There were three covariance models, and individual cluster characteristics identified for all datasets. All clusters were ellipsoidal, but some otherwise exhibited variation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cluster Count</th>
<th>Cluster ratios</th>
<th>Covariance Model</th>
<th>Cluster Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abortion - Miscarriage</td>
<td>5</td>
<td>25/70/64/36/41</td>
<td>VEV</td>
<td>ellipsoidal, equal shape</td>
</tr>
<tr>
<td>Congestive Heart Failure – Pulmonary Edema</td>
<td>2</td>
<td>124/116</td>
<td>VVV</td>
<td>ellipsoidal, varying volume, shape, and orientation</td>
</tr>
<tr>
<td>Delusion – Schizophrenia</td>
<td>2</td>
<td>120/120</td>
<td>VVV</td>
<td>ellipsoidal, varying volume, shape, and orientation</td>
</tr>
<tr>
<td>Dementia – Delirium</td>
<td>2</td>
<td>113/127</td>
<td>VVV</td>
<td>ellipsoidal, varying volume, shape, and orientation</td>
</tr>
<tr>
<td>Heart - Myocardium</td>
<td>2</td>
<td>120/120</td>
<td>VVE</td>
<td>ellipsoidal, equal orientation</td>
</tr>
<tr>
<td>Stroke - Infarct</td>
<td>2</td>
<td>121/119</td>
<td>VEV</td>
<td>ellipsoidal, equal shape</td>
</tr>
</tbody>
</table>

Table 2. Clustering results.
Two-dimensional cluster plot images of concept instances and predicates further illustrate cluster count and characteristics (Figure 1):

Abortion-Miscarriage  CHF-Pulmonary Edema  Delusion-Schizophrenia

Dementia-Delirium  Heart-Myocardium  Stroke-Infarct

Figure 1. Cluster plots of concept instances and predicates

For the five pairs exhibiting 2 clusters, the cluster ratios match or closely follow the 50 / 50 ratio of data instances (i.e., 120 instances for each concept) with little overlap where the ratio deviates from an evenly divided split. The best model for the dataset for Abortion and Miscarriage contained five clusters of varying proportions.

When we restored the labels to the data, and subjected it to an additional predictive process implementing Eigenvalue Decomposition Discriminant Analysis\(^3\^\) where each known class is modeled with the same covariance for each class, the results changed. The same covariance models resulting from the unlabeled clustering exercise were applied. The purpose of this was to observe which labels were associated with which cluster in each pair, when matched against possible classifications, although other outcomes from this modified process are also of interest. Abortion and Miscarriage are now represented by two clusters (because there are two classes represented in the data), and there is slightly reduced overlap for two of the concept pairs (see Figure 2).
While providing a basic, popular package for neural networks, nnet does not support multiple hidden layers. Experimentation revealed that five neurons in the single hidden layer produced the best results. Other limited experimentations enabled us to do some basic tuning to the other types of classifiers using the parameters already noted.

Each classifier yielded different composite accuracy scores, measured by the percentage of classifications in the testing data correctly predicted (Table 3). All but the Naïve Bayes classifier’s accuracy exceeded 80%.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.89</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.83</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.85</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 3. Classifier Accuracy

Precision, recall, and F1 scores for classifying individual concepts are indicated in Table 4.
<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Random Forest</th>
<th>Neural Network</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>Miscarriage</td>
<td>0.90</td>
<td>0.70</td>
<td>0.79</td>
<td>0.84</td>
</tr>
<tr>
<td>Delirium</td>
<td>0.83</td>
<td>0.81</td>
<td>0.82</td>
<td>0.73</td>
</tr>
<tr>
<td>CHF</td>
<td>0.89</td>
<td>0.74</td>
<td>0.81</td>
<td>0.59</td>
</tr>
<tr>
<td>Delusion</td>
<td>0.74</td>
<td>0.82</td>
<td>0.78</td>
<td>0.67</td>
</tr>
<tr>
<td>Pulmonary Edema</td>
<td>0.63</td>
<td>0.77</td>
<td>0.69</td>
<td>0.63</td>
</tr>
<tr>
<td>Abortion</td>
<td>0.90</td>
<td>0.93</td>
<td>0.91</td>
<td>0.87</td>
</tr>
<tr>
<td>Infarct</td>
<td>0.92</td>
<td>1</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>Myocardium</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.93</td>
</tr>
<tr>
<td>Dementia</td>
<td>0.91</td>
<td>1</td>
<td>0.95</td>
<td>0.91</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>1</td>
<td>0.95</td>
<td>0.97</td>
<td>1</td>
</tr>
<tr>
<td>Stroke</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>Heart</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4. Precision (P), recall (R), and F1 scores from classifier performance for individual concepts.

Discussion

This exercise demonstrates that people communicate differently about similar topics, even if (as in the case of Dementia and Delirium) that communication may share a substantial portion of contextual semantic elements. That difference can be found in analyzing data that captures how semantic elements interact.

In the first clustering exercise, 80% of the clustered concept pair datasets gave an affirmative answer to the study’s first question “Do unlabeled datasets of semantically related concept pairs exhibit two distinct clusters of equal or similar proportions when subjected to BIC modeling and EM clustering?”. Five of the six pairs identified two distinct clusters of proportions matching or similar to the count of instances for each concept. The cluster not exhibiting this behavior contained data for the Abortion-Miscarriage data. As mentioned earlier, the abortion data contained 116 instances, (as compared to 120 instances for all other concepts). The effect of this on the results is unknown, although the difference in the instance counts is quite small. When restricted to two clusters (there being two classes), modeling with the same covariance matrix, the Abortion-Miscarriage data behaved similarly to the other pairs.

Overall, the classifier training and testing exercise provided an affirmative answer to the question “Could a dataset for several concepts, where there is matching data for another semantically related concept for each concept in the dataset, be applied to training several different types of classifiers?”. Training multiple types of classifiers provides potential expectation of results for similar exercises using this type of data. The Support Vector Machine yielded the best overall accuracy. With the exception of Pulmonary Edema, F1 scores produced by the SVM for concept classification match or exceed those of the other types of classifiers, including delirium (0.82) and dementia (0.95). The Naïve Bayes classifier provided the overall lowest accuracy and F1 scores.

We compared these results to those of a similar task, WSD, where the researchers applied a new Expectation Maximization-based algorithm to determine the exact sense of the term “cold”\(^3\). Their top-performing model achieved 0.89 accuracy, which was comparable to the SVM we trained and most of the other state-of-the-art systems described in their work (the highest achieved being 0.93 by a Naïve Bayes application). In our similar task, we differentiated the sense of several terms, with the SVM, achieving F1 scores ranging from 0.69 to 1, with an average score of 0.89.
Implications in Future Work

Understanding differences in how individuals communicate about dementia and delirium may shed further light on distinguishing the two terms in text. To understand the results further, we reviewed instances of feature data for the 10 years, and found there is more variety in the predicate-opposite argument combinations in delirium as compared to dementia. Combinations that individually occur less than ten times comprise 37.7% of the delirium data, as compared to 18% for dementia. This means that, in the biomedical literature, authors communicate about delirium in more ways than dementia in terms of predicates and individual argument combinations. For predicate-semantic type combinations, there was much less variation, whereas combinations occurring less than ten times accounted for 1.4% and 4.5% of dementia’s and delirium’s data, respectively, suggesting that authors are more likely to communicate in more established predicate-semantic type patterns. These type of communication differences likely affected data behavior in the clustering and classifying exercises.

The data were drawn from the biomedical literature. Would similar data from clinical text yield the same results? There have been efforts to extract relational data from clinical text\textsuperscript{31, 32}. A proficient software application that extracts semantic predications from text may assist in answering this question. We are currently developing an application that extracts semantic predications from clinical text. Data with definitive instances of events such as delirium and dementia could be used to extract the initial semantic predications, to build a database of these data structures drawn from clinical text. This could be accomplished by using clinical text with a high certainty of such events.

A variation of sense differentiation in clinical text could address content where the author is uncertain of the event that is taking place, or if the resulting clinical text suggests this. For example, if a clinical note implied uncertainty regarding whether a patient had dementia or delirium, perhaps it could be determined by the way the author recorded it. Semantic predications extracted from clinical text where there is a high certainty of the events as recorded (as noted before) could be used to train an SVM classifier that in turn could be used to analyze data drawn from clinical text exhibiting less certainty. For the semantic predications extracted from the latter text, labels addressing the concepts of interest would have a generic indication, such as “Condition X”. For example, instances of dementia and delirium, where certainty of either is small, could be relabeled in this manner. Such work could reinforce an extended viewpoint on how word sense is identified, by taking into consideration the certainty with which a diagnosis or other assertion is recorded. It could complement and possibly incorporate research addressing the subjective properties of assertions in text\textsuperscript{33}.

A specific application could assess an author’s certainty in documenting diagnoses and other narratives addressing biomedical phenomena. For example, an application could alert a clinician discussing a diagnosis of Disease X that he or she is communicating about it in a manner consistent with Disease Y. This could be especially useful for diagnostically problematic conditions such as delirium. Application performance could be determined by processing relevant text, and interviewing the author to assess certainty in the recorded phenomenon.

In this study, we used semantic interaction rate data drawn from semantic predications to explicate the different senses of several semantically related pairs. There may be other NLP clinical text applications for which this type of data may be useful. It could drive temporal information retrieval\textsuperscript{34} by identifying documents with a high concentration of a given concept and frequent interactions with other semantic elements within a timeframe of interest. It could also be implemented in semi-supervised learning tasks, requiring temporal concept data, such as certain approaches to coreference resolution\textsuperscript{35}.

Limitations

The classifier exercise utilized a dataset limited to 12 concepts. However, it was intended to be an exploratory exercise implementing data of semantically related concepts, which was accomplished with this dataset.

This is an elementary exploration of analyzing interactions among semantic elements in text. We used a simple form of frequency data to capture semantic element interaction characteristics. It included temporal counts, but not other artifacts, such as, the most common predicate (e.g., PREVENTS) or opposite argument (e.g., calcium channel blocker), or other element for the given time period. Different expressions of semantic element interaction data may provide deeper insight to this phenomenon in text.

Factoring in SemRep’s performance in accurately identifying semantic relationships in the original text is outside of the scope of this paper, but would be an interesting subject for a future study addressing that topic.
Acknowledgements

We wish to thank Guy Divita for his contributions to this paper. This work was funded by the U.S. Department of Veterans Affairs, Health Services Research & Development (HSR&D), Project ID: CRE 12-321. This work was also supported in part by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

Conclusion

We applied interaction rate data extracted from semantic predications to clustering and classification exercises to determine if this data displayed expected clustering behavior, and could be used to classify concept instances where there were instances of several semantically related concept pairs in the dataset. Clustering of unlabeled data demonstrated two clusters of equal or similar proportions for five of the six concept pairs. When labels were included, all concept pairs demonstrated two clusters of equal or similar proportions. Four different classifiers were trained and tested. Accuracy for all classifiers ranged from 0.72 to 0.89, with a Support Vector Machine achieving the highest results. The outcomes for all these exercises pose interesting questions and potential applications in sense differentiation and other NLP tasks.

References

Learning Clinical Workflows to Identify Subgroups of Heart Failure Patients

Chao Yan, MS1, You Chen, PhD1, Bo Li, PhD1, David Liebovitz, MD2, Bradley Malin, PhD1
1Vanderbilt University, Nashville, TN; 2University of Chicago, Chicago, IL

Abstract
Heart Failure (HF) is one of the most common indications for readmission to the hospital among elderly patients. This is due to the progressive nature of the disease, as well as its association with complex comorbidities (e.g., anemia, chronic kidney disease, chronic obstructive pulmonary disease, hyper- and hypothyroidism), which contribute to increased morbidity and mortality, as well as a reduced quality of life. Healthcare organizations (HCOs) have established diverse treatment plans for HF patients, but such routines are not always formalized and may, in fact, arise organically as a patient’s management evolves over time. This investigation was motivated by the hypothesis that patients associated with a certain subgroup of HF should follow a similar workflow that, once made explicit, could be leveraged by an HCO to more effectively allocate resources and manage HF patients. Thus, in this paper, we introduce a method to identify subgroups of HF through a similarity analysis of event sequences documented in the clinical setting. Specifically, we 1) structure event sequences for HF patients based on the patterns of electronic medical record (EMR) system utilization, 2) identify subgroups of HF patients by applying a k-means clustering algorithm on utilization patterns, 3) learn clinical workflows for each subgroup, and 4) label each subgroup with diagnosis and procedure codes that are distinguishing in the set of all subgroups. To demonstrate its potential, we applied our method to EMR event logs for 785 HF inpatient stays over a 4 month period at a large academic medical center. Our method identified 8 subgroups of HF, each of which was found to associate with a canonical workflow inferred through an inductive mining algorithm. Each subgroup was further confirmed to be affiliated with specific comorbidities, such as hyperthyroidism and hypothyroidism.

Introduction
Heart failure (HF) is one of the most common indications for admission to the hospital among older adults. HF manifests in a clinically detectable manner when the heart is unable to supply an adequate flow of blood to meet the body’s needs. HF is an important contributor to both the burden and cost of national healthcare expenditures. Over five million people in the United States are estimated to exhibit HF to some degree and management of the disease costs the nation an estimated $32 billion annually. In 2001, the American Heart Association and American College of Cardiology refined the HF phenotype into four gross stages, which has led to the development and deployment of a wide array of management options for HF treatments. Yet, management of the disease is complicated by the fact that it often associates with a diverse collection of comorbidities (e.g., anemia, type 2 diabetes, various infections, and thyroid problems), which can manifest in different manners and combinations across the evolving stages of the disease. As a consequence, HF is also one of the conditions leading to high readmission rates in hospitals. Healthcare organizations (HCOs) have established treatment protocols and workflows for HF patients with different comorbidities and stages of progression. Additionally, to improve the definition and management of HF, various investigations have been conducted to computationally specify the phenotype and workflows affiliated with its management.

Traditionally, research has aimed to refine HF into several clinical subphenotypes based on heart-related issues, such as systolic or diastolic heart failure. These are natural subtypes that HCOs can rely upon to design specialized treatment plans or clinical workflows. While such research, and subsequent clinical designations, can assist HCOs to more effectively manage HF patients, they often rely on an expert informed perspective and experience. As a consequence, they involve a substantial amount of human effort and focus on clinical phenomena that are expected to categorize the HF population. To reduce human effort and learn clinical concepts (or management pathways) that are not necessarily anticipated, several studies have shown that data-informed methodologies can be invoked to infer complex comorbidities, clinical workflows, and care teams. Many of these studies rely on the co-accesses to patients’ electronic medical records (EMRs) to infer collaborative care teams or workflows for specific diseases, however, such studies have focused on all possible diseases and the workings of an HCO in general. In doing so, they have neglected how such views are influenced by conditioning the investigation on a specific complex disease, such as HF.

The investigation communicated in this paper is motivated by the expectation that complex diseases, like HF, are associated with a range of workflows in an HCO. These workflows are unlikely to be explicitly documented because
they are associated with subtypes of the disease and/or comorbidities that lead to *ad hoc* coordination. If such workflows can be detected through a data-informed method, they may be refined and resourced by an HCO to more effectively manage patients of a certain HF subtype.

Thus, in this paper we study four-months of EMR data, collected in 2010, from Northwestern Memorial Hospital for over 750 HF inpatient episodes. We introduce a data-informed framework to infer the underlying workflows that transpire in the clinical enterprise. We then map the learned workflows into a similarity measure to characterize patients into different subgroups. Finally, we show that these subgroups have a strong correlation with a range of diagnoses (e.g., hyper- and hypothyroidism) and procedures. Specially, our investigation suggests there are a minimum of 8 subgroups of HF patients, each of which is associated with a canonical workflow. It should be recognized from the outset that a subgroup does not indicate that they are a distinct population in their phenotype per se, but that they are a subgroup in the manner by which they are managed.

**Background**

This paper introduces a framework to identify subgroups of HF via inferred clinical workflows. Since this work involves workflow subgroup identification via inference methods, we take a moment to review related work in 1) workflow modeling and 2) subgroup discovery. When possible, we show how these methods have specifically been applied to HF populations.

**Workflow Modeling**

Workflow modeling and analysis has shown promise in a wide array of settings, ranging from general business management to specific clinical domain domains. For instance, van der Aalst and colleagues demonstrated how high-level Petri nets can model the workflows in an office environment, with a particular focus on how information systems support the control of office work\(^\text{17}\). They developed a workflow management system (WMS) prototype based on such formalizations. They subsequently extended the notion of a WMS to support dynamic changes\(^\text{18}\). Workflow modeling from a clinical perspective is more complex than many office settings because HCOs are composed of a large number of interacting departments and individuals who coordinate as availability and need dictates.

Still, there has been some success in this domain. In particular, Chen and colleagues introduced a method to infer clinical workflows and measure their efficiency via the utilization of EMR event logs. It was shown that these workflows naturally partition into four general types according to their average and variance in their efficiency\(^\text{11}\). They posited that certain inefficiencies were likely due to the complexity of the patients. While the methods introduced in their investigation enabled the evaluation of workflow efficiency, it did not condition the workflows on specific patient phenotypes or determine if subgroups for the management of a specific disorder led to the manifestation of disparate workflows. EMR access logs were also used by Li and colleagues to infer workflows through a method based on hidden Markov models (HMMs)\(^\text{19}\). These HMMs were utilized to characterize the behavior of EMR users, as well as detect anomalous activities. However, this investigation did not study the clinical meaning of the workflow or how they could be specialized to certain patient subgroups.

**Subgroup Identification**

There is evidence to suggest that identifying subgroups of patients based on their clinical conditions can be applied to design personalized treatment plans. For instance, Mugge and colleagues identified a subgroup of patients with nonrheumatic atrial fibrillation (Afib) with an increased risk for cardiogenic embolism by assessing left atrial appendage function\(^\text{20}\). They identified two distinct patient groups according to appendage flow patterns: 1) well-defined peak filling with visible fibrillatory contractions of the appendage wall and 2) irregular, very low, peak filling with almost no visible appendage contractions. While such subgroup identification is notable from a descriptive perspective, it focuses on clinical diagnosis and, thus, neglects how to design management routines for Afib patients in a healthcare environment that is subject to communication challenges and resource limitations. Soulakis and colleagues\(^\text{16}\) utilized the EMR access logs of over 500 HF patients to identify seven networks of around 5000 care providers. However, this work is limited in that it neglects the association between the network of care providers and the clinical conditions of the patients.

**Methods**

We designed a data-informed framework to identify HF subgroups to consist of three steps: 1) infer patient subgroups through event sequences, 2) learn a workflow for each subgroup and 3) assign diagnosis and procedure codes as phenotype labels for each subgroup. To gain intuition into how this framework works, we begin with an introduction of the dataset used in this investigation.
**Dataset**

The dataset for this study is summarized in Table 1. The records are drawn from comprehensive access logs and billing information derived from the EMR system at Northwestern Memorial Hospital (NMH) over a four month period during 2010. This dataset, which we refer to as NMH-D, consists of 1,138,555 access events distributed over 16,567 unique inpatient stays. We extracted all patients diagnosed with heart failure by selecting any patient with an ICD-9 code in the range 428.0 through 428.9 and we refer to this specific HF extract as the NMH-HF dataset. Each access event is affiliated with the following attributes: 1) pseudonym of the inpatient, 2) ID of the EMR user, 3) reason for the access event (as designated by the user according to a pull-down list), 4) date and timestamp for the access, 5) general physical location in NMH where the patient is located, and 6) the clinical service on which the patient is managed (e.g., general medicine vs. obstetrics). Each inpatient episode is affiliated with its ICD-9 billing codes, which were assigned after discharge.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accesses</th>
<th>Patients</th>
<th>Reasons</th>
<th>Number of 2-blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMH-D</td>
<td>1,138,555</td>
<td>16,567</td>
<td>142</td>
<td>(not computed)</td>
</tr>
<tr>
<td>NMH-HF</td>
<td>272,685</td>
<td>785</td>
<td>126</td>
<td>5823</td>
</tr>
</tbody>
</table>

Table 1. A summary of the datasets in this study.

We use the *Reason for Access* in the dataset, as opposed to the EMR user, as the smallest level of granularity associated with a workflow to mitigate noise in the analysis. Specifically, for each inpatient *i*, the corresponding event sequence *R* defined in Table 2 is a series of ordered reasons for access. To orient the reader, the following is an example of a sequence of reasons one might encounter for a certain inpatient:

... → Attending Physician/Provider → Primary Staff Nurse → Resident-Inpatient Primary Service → Pharmacist → ...

In this work, we define an *n*-block as a sequence of *n* consecutive access events (with *n* ≥ 2). Thus, a 2-block is defined as two consecutive access events. For instance, if an access event *Attending Physician/Provider* has a consecutive access event *Primary Staff Nurse*, then *Attending Physician/Provider* → *Primary Staff Nurse* is the corresponding 2-block. In fact, 2-blocks reflect the order relations between two neighbor events. If 2-blocks appear infrequently in event sequences, then the order relation between its containing two events are weak, which could be utilized by our framework to filter weak order relations, most of which are noise in EMR system functionality.

**Subgroups Identification Framework**

The framework is composed of three main components: 1) generate patient subgroups via refined event sequences, 2) learn workflows for the subgroup population, and 3) extract diagnosis and procedure labels for subgroups. Here,
refined event sequences are the transformation of original event sequences through applying filtering based on frequency. To help the reader understand our framework, we provide a legend of the common notation invoked throughout this paper. The specific algorithmic process associated with the framework is communicated in Figure 2.

Table 2. A legend of the notation used in this paper.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>A set of event sequences, where an event sequence is affiliated with one patient episode</td>
</tr>
<tr>
<td>$D_i$</td>
<td>A set of event sequences filtering high frequency reasons in $D$</td>
</tr>
<tr>
<td>$D'$</td>
<td>A set of refined event sequences</td>
</tr>
<tr>
<td>$\phi$</td>
<td>A tuple set of reasons and corresponding frequency</td>
</tr>
<tr>
<td>$B_{(2_block)}$</td>
<td>A tuple set of 2-block and corresponding frequency</td>
</tr>
<tr>
<td>$B_{(n_block)}$</td>
<td>A tuple set of blocks and corresponding frequency</td>
</tr>
<tr>
<td>$R^i$</td>
<td>An event sequence in $D$ charactering inpatient $i$’s episode</td>
</tr>
<tr>
<td>$R'^i$</td>
<td>A refined event sequence in $D'$ characterizing inpatient $i$’s episode</td>
</tr>
<tr>
<td>$r_t$</td>
<td>A reason appearing at time stamp $t$</td>
</tr>
<tr>
<td>$b_{i+1}$</td>
<td>A 2-block with time stamp $t$ and $t+1$</td>
</tr>
<tr>
<td>$D_{(2_block)}$</td>
<td>A 2-block set extracted from $D_i$</td>
</tr>
<tr>
<td>$D'_{(2_block)}$</td>
<td>A refined 2-block set from $D_{(2_block)}$ by filtering lowest frequency 2-blocks</td>
</tr>
<tr>
<td>$D_{(n_block)}$</td>
<td>A set of blocks extracted from $D'$, where each block can have a different length</td>
</tr>
<tr>
<td>$D'_{(n_block)}$</td>
<td>A set of blocks obtained by filtering lowest frequency blocks in $D_{(n_block)}$</td>
</tr>
<tr>
<td>$m_r$</td>
<td>The number of filtered highest frequency reasons in $\phi$</td>
</tr>
<tr>
<td>$m_b$</td>
<td>The number of filtered lowest frequency 2-blocks in $D_{(2_block)}$</td>
</tr>
<tr>
<td>$m_g$</td>
<td>The number of filtered lowest frequency blocks in $D_{(n_block)}$</td>
</tr>
<tr>
<td>$M_{PB}$</td>
<td>A binary matrix characterizing the relationship between patient episodes and blocks in $D'_{(n_block)}$</td>
</tr>
<tr>
<td>$M_{PB}'$</td>
<td>A matrix transformed from $M_{PB}$ using a polynomial kernel</td>
</tr>
</tbody>
</table>

1) Generate Patient Subgroups by Refined Access Sequences

The subgroup identification process is partitioned into five steps: 1) filter high frequency reasons contained in each event sequence in $D$ into $D_i$, 2) generate 2-blocks in $D_i$ and create a set $D_{(2\_block)}$, 3) filter low frequency 2-blocks in $D_{(2\_block)}$ into $D'_{(2\_block)}$, 4) use 2-blocks in $D'_{(2\_block)}$ to refine sequences in $D_i$ to form a new sequence set $D'$ and 5) clustering subgroup patients by using similarity of block elements in refined sequences in $D'$. Each of these steps is detailed in the following descriptions:

Filter high frequency reasons. Each raw event sequence in $D$ can contain high frequency event reasons. We remove high frequency reasons because they correspond to the most general aspects of the workflow. These are unlikely to communicate clinical context that is critical to modeling a specific workflow. For example, both Primary Staff Nurse and Primary Assistive Staff appear the greatest number of times in event sequences. While it is anticipated that nurses provide support to patient care, they are critical to almost all aspects of the inpatient setting. In many respects, general nursing staff is akin to the stop words (e.g., prepositions or articles) in natural language text. And, as many investigations in natural language processing have illustrated, such information, can be triaged to improve pattern discovery. The new event sequence set $D_i$ is generated by filtering the high frequent reasons.

Generate 2-blocks and filter low frequency (or noise). These two steps are incorporated because the event sequences are extracted from the access logs of EMR system, which contain noise in the order of relations. We assume that noisy relations exist in low frequency blocks and, thus, filter weak relations from the event sequences $D_i$.

To do so, each event sequence in $D_i$ is segmented into blocks by invoking the 2-blocks in $D'_{(2\_block)}$. Thus, each event sequence is represented by varying sized blocks. With this transformation and linkage, $D_i$ is transformed into $D'$. Thus, $D'$ is the set of sequences where high frequent reasons and low frequent 2-blocks are both removed. The
processing details are described from lines 6 through 17 in the algorithm in Figure 2. For example, after the process above, a patient in our dataset has a new sequence formed by two blocks $b_1 \rightarrow b_2$, where $b_1$ is a reason sequence:

```
Resident - Outpatient/ED/Proc Primary \rightarrow Patient Care \rightarrow Radiology Technologist \rightarrow Registration
```

and $b_2$ is

\[
\text{Rehab Assigned Therapist} \rightarrow \text{Consultant} \rightarrow \text{Rehab Assigned Therapist} \rightarrow \text{Charging/Orders} \rightarrow \text{Med Rec Coding}
\]

$D'$ is represented by linkage of new blocks and we extract all of the blocks in $D'$ into $D'(n\_block)$. At this point, we filter out low frequent blocks in $D'(n\_block)$ to generate a new set $D'(n\_block)$.

Next, we cluster patients into subgroups using their associated blocks in $D'(n\_block)$. To do so, we generate a patient-by-block matrix $M_{PB}$ to represent relations between patients and blocks. We then apply a polynomial kernel on $M_{PB}$ to transform it to a new matrix $M_{PB}'$ and perform $k$-means clustering.\(^1\)

---

**Input:** $D$, a set of event sequences; $m_r$, the number of filtered lowest frequency 2-blocks; $m_b$, the number of filtered lowest frequency blocks; $m_b$, the number of filtered lowest frequency blocks; $m_g$, the number of filtered highest frequency blocks; $m_r$, the number of filtered highest frequency blocks;

**Output:** $C = \{C_1, C_2, \ldots, C_s\}$, subgroups of patients

```
1: Let $\phi = \{(\varphi_i, f_i)\}$ ← (reason ID, frequency) tuple set from $D$
2: $D_r \leftarrow D \setminus (\text{top } m_r \text{ reasons with high frequency in } \phi)$
3: $D'(2\_block) = \{(\varphi_i, f_i)\}$ ← 2-block set from $D_r$
4: Let $B(2\_block) = \{(\varphi_i, f_i)\}$ ← (2-block, frequency) tuple set from $D_r$
5: $D'(2\_block) \leftarrow D'(2\_block) \setminus (\text{top } m_b \text{ 2-blocks with low frequency in } B(2\_block))$

\[
\text{Output: } \phi = \{(\varphi_i, f_i)\} \leftarrow \text{reason ID, frequency) tuple set from } D
\]

\[
\text{Return: } C = \{C_1, C_2, \ldots, C_s\}, \text{ subgroups of patients}
\]

---

**Figure 2.** Pseudocode of the cluster generation algorithm.

2) **Learn workflows for each subgroup**

Each subgroup is clustered using blocks in $D'(n\_block)$. Each event sequence (a patient episode) is characterized by these blocks. We group all blocks characterizing a subgroup into a workflow to represent the clinical process for this type of HF patient. We then invoke an inductive mining algorithm (as implemented in ProM\(^{22}\)) to infer and visualize workflows.

3) **Extract Diagnosis and Procedure Labels**

---

\(^1\) Applying the kernel makes it easier to separate groups by projecting the data into a higher set of dimensions.
Care must be taken when learning workflows through a data-informed strategy, as they may not have labels that readily translate into administrative applications. This is important because the clinical workflows that are based on expert knowledge are associated with known semantics. Thus, we aim to relate inferred workflows to clinical context. To do so, we assign labels to each subgroup’s workflow by discovering the billing codes that are the most discriminative for the workflow.

The labels for each workflow are derived from a series of processes. First, we extract the most frequent diagnosis and procedure codes from all inpatients in each subgroup. Next, we apply a Z-test (hypothesis testing method based on proportionality) to compute the p-value for each diagnosis and procedure codes in each subgroup\(^23\). For each subgroup \(C_a\), the most distinguishable billing codes between pairs of subgroups \((C_a, C_x)\), \(x\neq a\), are extracted. Finally, we union these the distinguishable codes to characterize each subgroup and their affiliated workflows.

**Table 3. A summary of the number of patients and representative conditions for the HF subgroups.**

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Size</th>
<th>Billing Terms with High Frequency</th>
<th>Phenotype</th>
</tr>
</thead>
</table>
| \(C_1\)  | 447  | 401.9: Unspecified essential hypertension  
        416.8: Other chronic pulmonary heart diseases  
        425.4: Other primary cardiomyopathies  
        414.0: Coronary atherosclerosis of native coronary artery  
        V45.81: Aortocoronary bypass status  
        V58.61: Long-term (current) use of anticoagulants | General HF                                     |
| \(C_2\)  | 14   | 276.2: Acidosis  
        280.0: Iron deficiency anemia secondary to blood loss  
        578.9: Hemorrhage of gastrointestinal tract  
        V65.3: Dietary surveillance and counseling  
        V10.04: Personal history of malignant neoplasm of stomach | HF associated with hemorrhage                  |
| \(C_3\)  | 15   | 428.4: Combined systolic and diastolic heart failure  
        276.8: Hypopotassemia  
        V15.05: Allergy to other foods  
        V15.06: Allergy to insects and arachnids | HF associated with anaphylactic reaction        |
| \(C_4\)  | 22   | 263.9: Unspecified protein-calorie malnutrition  
        038.9: Unspecified septicemia  
        584.9: Acute kidney failure  
        V45.11: Renal dialysis status | HF associated with renal failure and sepsis    |
| \(C_5\)  | 13   | 729.5: Pain in limb  
        276.5: Dehydration  
        275.4: Hypocalcemia  
        V42.0: Kidney replaced by transplant | HF associated with renal transplantation       |
| \(C_6\)  | 58   | 276.7: Hyperpotassemia  
        584.9: Acute kidney failure  
        275.3: Disorders of phosphorus metabolism  
        428.2: Acute on chronic systolic heart failure  
        287.5: Thrombocytopenia  
        V45.11: Renal dialysis status | Hyperthyroidism HF                             |
| \(C_7\)  | 5    | (too small to make a determination)                                                               | (too small to make a determination)            |
| \(C_8\)  | 6    | (too small to make a determination)                                                               | (too small to make a determination)            |
| \(C_9\)  | 34   | 244.9: Unspecified acquired hypothyroidism  
        733.0: Osteoporosis  
        401.9: Unspecified essential hypertension  
        425.4: Other primary cardiomyopathies | Hypothyroidism HF                             |
| \(C_{10}\) | 171  | 403.9: Hypertensive chronic kidney disease  
        585.9: Chronic kidney disease  
        584.9: Acute kidney failure  
        285.9: Anemia  
        V45.01: Cardiac pacemaker in situ  
        V45.82: Percutaneous transluminal coronary angioplasty status | Acute or chronic renal failure HF             |
Results
This section reports on results from 1) a general view of the 10 identified distinct subgroups and 2) a case study of two representative subgroups - in terms of their affiliated workflows and the diagnosis and procedure codes that distinguish the subgroup from others.

HF Subgroups Identified
By applying the framework to the 785 CHF inpatient episodes, we discovered 10 distinct subgroups as summarized in Table 3. For reference purposes, we use $C_i$ to represent subgroup $i$. In this table, each subgroup is labeled using descriptions of the representative diagnosis and procedure billing codes. For instance, $C_1$ is a subgroup affiliated with general heart failure, $C_7$ is affiliated with allergies associated with HF, and $C_8$ is affiliated with hypothyroidism. We note that we neglect subgroups $C_7$ and $C_8$ because of their small size (only 5 and 6 patients, respectively). The representative billing codes come from two sources: 1) distinct codes between this subgroup and any other subgroup, and 2) high frequency codes associated with this subgroup.

Table 3 also reports the size and clinical label of each subgroup. It can be seen that each subgroup has a distinct specialized phenotype. For instance, $C_1$, the largest subgroup, is affiliated with the diagnosis of HF, which indicates that most of the patients went through a process associated with management of general comorbidities (e.g., hypertension, primary cardiomyopathies, and coronary atherosclerosis). $C_{10}$, the second largest subgroup, is affiliated with chronic kidney disease (CKD).

Case Study of Hyperthyroidism and Hypothyroidism in HF
To understand the intuitive nature of the identified subgroups, we report on a case study that compares two subgroups. Specifically, we focus on subgroups $C_6$ and $C_9$, which are similar in the number of patients they cover (58 and 34, respectively) and correspond to two typical HF subtypes. We illustrate the differences in these subgroups in terms of their clinical concepts and inferred workflows.

Clinical Differences. Figure 3 shows the concordance of the frequency distribution for the most significant codes (based on their p-values) affiliated with the $C_6$ and $C_9$ subgroups. Specifically, each ($x,y$) point corresponds to a specific code, where $x$ and $y$ is the proportion of patients in $C_6$ and $C_9$ who received the code, respectively. As such, codes that are close to the dashed diagonal line indicates they have a similar frequency in the two investigated subgroups. Clearly, these codes do not distinguish between the subgroups. By contrast, codes that are distant from the line can distinguish one subgroup from the other. In this figure, we marked diagnosis codes with a star and procedure codes as a circle.

We found that the patients in cluster $C_6$, were primarily diagnosed with i) hyperpotassemia, ii) disorders of phosphorus metabolism, iii) acute kidney failure, unspecified, and iv) thrombocytopenia. This combination of diagnoses makes sense intuitively. This is because excess potassium and phosphorus caused by hyperthyroidism are both associated with kidney failure\textsuperscript{21}. Based on knowledge of these symptoms, it can be inferred that this subgroup suffers from Hyperthyroidism HF.

By contrast, the 34 patients in cluster $C_9$ have the following diagnosis labels: i) unspecified acquired hypothyroidism, ii) osteoporosis, and iii) unspecified essential hypertension. Osteoporosis is obviously associated with hypothyroidism because of a decreasing amount of calcium. And thus, we label $C_9$ with Hypothyroidism HF.

These HF subgroups demonstrate that differences in the inferred learned workflows may be associated with the clinical status, as well as procedures performed on the patients.

Workflows Differences. Though subgroup $C_6$ covers a larger number of patients than $C_9$, we find that it exhibits a simpler workflow structure (as shown in Figure 4). To illustrate, we first compare the main reasons of these two workflows. Subgroup $C_6$ contains: i) Other Physician, ii) Covering Therapist, iii) Coordinator, and iv) Advanced Practice Nurse as the main reasons (and corresponding roles), which appear to be affiliated with a generic healthcare process. By contrast, the workflow for $C_9$ has a more complicated structure, in which there is a greater diversity in the reasons. Specifically, this structure includes i) Student Nurse, ii) Consultant, iii) Patient Care, iv) Assigned Staff, v) Advanced Practice Nurse, vi) Coordinator, vii) Anesthesiologist, viii) Dietary Clerk, ix) Radiologist, and x) Rehab Assigned Therapist.

Moreover, this workflow contains some special reasons, such as xi) Radiology Nurse/Resident/Technologist and xii) Radiology management, which are likely linked with the diagnosis of osteoporosis induced by hypothyroidism. This case study suggests that our initial hypothesis – that HF subtypes associate with different workflows and management
processes – has standing. Moreover, we believe this case study is a clear illustration of how data-informed workflow mining can be leveraged to learn different subgroups of patients that can be subsequently labeled in a clinically meaningful manner.

Figure 3. Concordance in the frequency of the diagnosis and procedural billing codes between HF subgroups C_6 and C_9.

Figure 4. A simplified view on the workflow structure of subgroups (a) C_6 and (b) C_9.
Discussion

We introduced a method to identify subgroups of HF through refined event sequences and subsequently infer workflow and phenotype for every subgroup. This approach is substantially different from the traditional definition of phenotypes rooted solely on metabolic and clinical observations. The subgroups identified in our framework share similar workflow patterns, suggesting they are managed in a similar way in clinical practice. To the best of our knowledge, this is the first investigation to show that subgroups of a complex disorder, such as HF, can be learned through workflows (in the form of event sequences) in the clinical enterprise. Our findings are further notable because they suggest that workflows stretching across departments and wards of an HCO can be learned from EMR utilization. We believe this methodology provides opportunities to make management strategies explicit and tune resource allocations accordingly. At the same time, we believe the learned workflows have standing because they were shown to correlate with diagnoses and procedures codes exhibited in the corresponding patient groups (information that was not used in the clustering process). As such, it may be possible to develop predictive models that assign a patient to a predefined and semi-personalized management regimen.

Despite our discoveries, we acknowledge that this is a pilot study on an HF population. There are several limitations of this project, which we wish to highlight for further refinement and future investigation. First, we relied on ICD-9 codes in the 428.* range, rather than rigorously validated computational phenotypes of HF to define our cohort. Though the HF phenotype is considered a relatively well-defined diagnosis, it is a disease with multiple stages and confounding factors (as our hyperthyroidism and hypothyroidism subgroups illustrated). It is further conceivable that some of the patients who presented to Northwestern Memorial Hospital for a certain primary diagnosis (e.g., stroke) might have been treated for HF without its documentation in an ICD-9 billing code. As such, we believe that our selection criteria enable a highly precise investigation, but does not cover the gamut of HF patients.

Second, the workflows associated with the HF subgroups were only reviewed by one clinician. Rather, we mainly relied on identifying subgroups of HF in an information theoretic sense (e.g., similarity analysis of sequences of access reasons). Our method identified 10 subgroups and it appeared as though 8 had a clear clinical context differentiating from other subgroups (while 2 were too underpowered due to a small number of patients to make any judgement about). Still, before inferred workflows can be relied upon, they will require review by additional administrative and clinical experts to determine if they can be translated into decision support tools for an HCO.

Finally, we recognize that the size of the HF cohort is relatively small. During the four months of documented inpatient stays, we encountered less than a thousand patients treated for HF. Although our work yielded meaningful findings, we may not have captured all of the notable workflows or diagnostic labels for such workflows. For such an investigation to be useful for the HCO, we will need to investigate a large sample size over a longer period of time. It would also be ideal to compare the workflows, and the associated labels, with EMR data from other healthcare systems.

Conclusions

HF is a complex condition accompanied by diverse complications that progress across, a minimum of, four stages. HCOs have adopted various strategies (e.g., optional treatment plans for different developmental stages and complications) to improve quality of life for HF patients and reduce burden, as well as cost, for healthcare systems. Identifying subgroups of HF can assist in the management of patients with this disease. We introduced a data-informed framework to identify subgroups of HF patients through utilization of the EMR. For each subgroup, we provided external validation via the diagnosis and procedure codes of the corresponding patients. Our framework was evaluated on the event sequences of 785 HF inpatients from a large academic medical center. In doing so, we identified 8 HF subgroups, each of which was confirmed to be associated with a specific condition of HF (e.g., hyperthyroidism and hypothyroidism). Furthermore, each subgroup was characterized by different patterns of workflow. For instance, hyperthyroidism associated HF involved more complex workflows than hyperthyroidism HF. We acknowledge that this investigation is a pilot study and further investigation is required with administrative review and validation across disparate healthcare enterprises.

Acknowledgements

This research was supported, in part, by the following grants from the National Library of Medicine at the National Institutes of Health K99LM011933, R00LM011933 and R01LM010207. The content in this work is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.
References

Breach Risk Magnitude: A Quantitative Measure of Database Security

William A. Yasnoff, MD, PhD1,2
1NHII Advisors, Arlington, VA; 2Johns Hopkins University, Baltimore, MD

Abstract
A quantitative methodology is described that provides objective evaluation of the potential for health record system breaches. It assumes that breach risk increases with the number of potential records that could be exposed, while it decreases when more authentication steps are required for access. The breach risk magnitude (BRM) is the maximum value for any system user of the common logarithm of the number of accessible database records divided by the number of authentication steps needed to achieve such access. For a one million record relational database, the BRM varies from 5.52 to 6 depending on authentication protocols. For an alternative data architecture designed specifically to increase security by separately storing and encrypting each patient record, the BRM ranges from 1.3 to 2.6. While the BRM only provides a limited quantitative assessment of breach risk, it may be useful to objectively evaluate the security implications of alternative database organization approaches.

Introduction
Breaches involving unauthorized disclosure of substantial numbers of identified electronic medical records are occurring with increasing frequency. These events damage the trust of the public in health information systems, may seriously harm patients whose data is exposed, and are very costly to the organizations responsible for holding the data. As a result, preventing data breaches has become a very high priority in the design and implementation of clinical information systems.

This paper describes a quantitative assessment methodology to evaluate the impact of database architecture on the risk of security breaches. By applying this quantitative methodology, it is shown that alternative architectures can reduce these risks. This suggests that use of quantitative breach risk measures may be useful in guiding the design of more secure clinical information systems.

Need for Quantitative Breach Risk Assessment
Databases have traditionally organized information to facilitate rapid search and retrieval operations, while the security of the stored information has in general been a secondary consideration. Although it is relatively easy to measure and/or calculate the response time for database search operations, there are no existing quantitative measures that can assist information architects in evaluating the potential for security breaches when considering alternative possibilities for organizing and storing data.

There have been a number of prior efforts to develop security metrics. Zhang et al described a semi-quantitative approach for assessing enterprise security that attempts to model the behavior of potential attackers using the variables of intent, objective, and consequence as input to a Markov decision process. However, the specific parameters of the model are unknown and must be estimated in each case. Harel et al described a “misuseability weight” that evaluates the sensitivity level of exposed data. This sensitivity is used to calculate an “M-score” that incorporates both the quantity and quality of the information at risk, but does not include any assessment of the difficulty of obtaining access. Aissa et al defined a value-based cybersecurity measure that attempts to quantify the potential dollar losses of intrusions for specific stakeholders. Bhattacharjee et al proposed a more complex scheme for assessment of overall enterprise level information security risk that assigns a specific risk level to each enterprise information asset based on its “threat-vulnerability” pair, and suggests specific actions of decreasing stringency to ameliorate risks for high, medium, and low risk assets respectively. Finally, Aime et al described the risks of security metrics in general, noting that the use of such metrics does not replicate the performance of experts and therefore they must not be used in isolation. However, none of this prior work describes a metric that can provide specific, easily calculated results incorporating both the quantity and degree of difficulty of accessing information.

Consistent with prior recommendations to identify and establish measures that provide insight into potential risks, the breach risk magnitude measure was developed. Its fundamental assumption is that breach risk is related to the number of potential records that could be exposed by a specific sequence of authentication processes (typically passwords), while it is decreased when more such authentication steps are required for access. In other words, additional authentication steps make it less likely that an unauthorized access will occur. This is consistent with an
intuitive understanding of the incentives of a potential attacker: a target is more attractive when either fewer obstacles need to be overcome to gain access or more records can be obtained by circumventing those obstacles.

**Desiderata of Breach Risk Measure**

Before attempting to define any quantitative measure of a previously subjective characteristic, its desired properties should be considered and described. These properties, which represent the requirements that must be met, allow the characteristics of the resultant measure to be assessed to determine if the goals have been achieved. An alternative set of requirements might lead to a substantially different measurement approach.

To be useful, a breach risk measure should be:

1. Roughly proportional to our intuitive sense of the level of risk;
2. Higher when the level of risk increases and lower when it decreases;
3. Able to express a very wide range of risk levels;
4. Straightforward to calculate across a variety of systems; and
5. Easily understood and interpreted.

The first three properties represent accuracy, proportionality, and scalability respectively, while the last two relate to ease of use.

**Methods**

The development of the measure was guided by the intuitive notion that a system allowing access to a larger number of records with a smaller number of authentication steps is inherently more vulnerable to compromise. This is consistent with concepts of event likelihood and impact/severity described in the risk assessment guidelines from NIST\(^8\). In this case, the likelihood of an event is assumed to be inversely proportional to the number of authentication steps, while the impact/severity of an event relates to the number of records exposed.

Based on these ideas, the breach risk assessment is defined as the ratio of the number of records that are accessible divided by the number of authentication steps required to enable that access. The larger the value, the higher the risk. To accommodate a wide range of values, the breach risk magnitude (BRM) is expressed as the common logarithm of the breach risk assessment.

As an example, assume a system with one million records, all of which can be accessed by the system administrator with a single authentication step. In this case,

\[
\text{Number of accessible records} = 1,000,000 \\
\text{Number of authentication steps} = 1 \\
\text{Ratio of accessible records to authentication steps} = \frac{1,000,000}{1} = 1,000,000 \\
\text{Breach Risk Magnitude (BRM)} = \log_{10} (1,000,000) = 6
\]

By expressing the measure as a log value, much like the Richter scale for earthquake magnitude\(^9\), an extremely wide range of values can be communicated quickly and easily. For example, the above system with only 100,000 records would have a BRM of 5, while the same system with 10,000,000 records would have a BRM of 7. Thus, each increase of one in the BRM value represents 10 times the potential security vulnerability.

In a given system, there are typically several different classes of users that access records using different procedures. There may be individuals who only are permitted to access their own record, researchers who may access a large subset (or all) of the records, and a system administrator who typically has access to all the records. Recognizing that the security vulnerability of a system relates to its weakest link, the BRM value for a given system is calculated as the maximum value for any class of users with access to the stored records. In many such systems, the user class with the maximum BRM will be the system administrator.

Calculation of the BRM measure does not require access to operational database systems. The input parameters of number of records and number of authentication steps for each user are evident from the organization of a database implementation. For example, any user given the capability to search the entire database by definition has access to all the records.
Results of Breach Risk Magnitude (BRM) Application

The BRM measure was evaluated for several sample database system configurations, each with one million records. Table 1 shows the calculation for a standard relational database management system. Three types of users are allowed in this example: 1) individuals with access to only their own record; 2) searchers who may access all the records for search purposes; and 3) system administrators who may access all data for any purpose. The BRM value for this system is 6.

Table 2 shows the effect of requiring a second authentication step for searching or for the system administrator to access the records. While the risk ratio drops by 50%, the BRM measure only decreases to 5.7.

Table 3 shows the impact of separating the demographic data from the remainder of the data for each person’s record, a commonly proposed technique for increasing security. A separate password is required to access the demographic and non-demographic data. While the searcher users no longer have access to complete records, the system administrator can still access all the data, albeit with three passwords. Therefore, the BRM for this approach only decreases to 5.52.

Table 4 shows the results for a different data storage architecture known as the personal grid10. The personal grid is specifically designed to improve the security of personal information by storing each person’s data in a separate file with its own separate encryption. The encryption/decryption key for each file consists of two distinct and independent parts, one supplied by the user and the other by the system.

Since each record is separately encrypted and there is no inverted index of the records (like in a relational database), searching with the personal grid must be done sequentially one record at a time. To accelerate this prohibitively slow process, search operations are parallelized using cloud computing. When a search is needed, a large number of servers (e.g., between 500 and 10,000) are allocated temporarily (in a cloud computing environment) and each server simultaneously processes its share of the records. For example, with a database of 1,000,000 records, each search server would evaluate between 100 and 2,000 records (for 10,000 or 500 search processors respectively).

### Table 1. Breach Risk Magnitude calculation for Relational Database Management System (RDMS)

<table>
<thead>
<tr>
<th>User Type</th>
<th>Authentications Required</th>
<th># Records Accessed</th>
<th>Breach Risk</th>
<th>BRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Login (1)</td>
<td>1</td>
<td>1/1</td>
<td>0</td>
</tr>
<tr>
<td>Searcher</td>
<td>Login (1)</td>
<td>1 million</td>
<td>1 million/2</td>
<td>5.7</td>
</tr>
<tr>
<td>System Admin</td>
<td>Login (1)</td>
<td>1 million</td>
<td>1 million/2</td>
<td>5.7</td>
</tr>
<tr>
<td>OVERALL</td>
<td>Max BRM</td>
<td></td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

### Table 2. Breach Risk Magnitude calculation for RDMS requiring double authentication for access to entire dataset

<table>
<thead>
<tr>
<th>User Type</th>
<th>Authentications Required</th>
<th># Records Accessed</th>
<th>Breach Risk</th>
<th>BRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Login (1)</td>
<td>1</td>
<td>1/1</td>
<td>0</td>
</tr>
<tr>
<td>Searcher</td>
<td>Login and Entire dataset (access (2)</td>
<td>1 million</td>
<td>1 million/2</td>
<td>5.7</td>
</tr>
<tr>
<td>System Admin</td>
<td>Login and Entire dataset (access (2)</td>
<td>1 million</td>
<td>1 million/2</td>
<td>5.7</td>
</tr>
<tr>
<td>OVERALL</td>
<td>Max BRM</td>
<td></td>
<td></td>
<td>5.7</td>
</tr>
</tbody>
</table>

### Table 3. Breach Risk Magnitude calculation for RDMS requiring separate authentication for access to demographic and non-demographic records

<table>
<thead>
<tr>
<th>User Type</th>
<th>Authentications Required</th>
<th># Records Accessed</th>
<th>Breach Risk</th>
<th>BRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Login (1)</td>
<td>1</td>
<td>1/1</td>
<td>0</td>
</tr>
<tr>
<td>Searcher</td>
<td>Login and Access to de-identified non-demographic records (2)</td>
<td>No complete records</td>
<td>0/2</td>
<td>n/a</td>
</tr>
<tr>
<td>System Admin</td>
<td>Login, Access to de-identified non-demographic records, Access to demographic records (3)</td>
<td>1 million</td>
<td>1 million/3</td>
<td>5.52</td>
</tr>
<tr>
<td>OVERALL</td>
<td>Max BRM</td>
<td></td>
<td></td>
<td>5.52</td>
</tr>
</tbody>
</table>

### Table 4. Breach Risk Magnitude calculation for personal grid architecture

<table>
<thead>
<tr>
<th>User Type</th>
<th>Authentications Required</th>
<th># Records Accessed</th>
<th>Breach Risk</th>
<th>BRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Login; System master password; System record password (3)</td>
<td>1</td>
<td>1/3</td>
<td>-0.48</td>
</tr>
<tr>
<td>Search Server</td>
<td>Login (3 passwords); System master password; System record password (5)</td>
<td>100 (10,000 servers)</td>
<td>100/5</td>
<td>1.3</td>
</tr>
<tr>
<td>System Admin</td>
<td>Login; Cannot access individual records</td>
<td>0</td>
<td>0/1</td>
<td>n/a</td>
</tr>
<tr>
<td>OVERALL</td>
<td>Max BRM</td>
<td></td>
<td></td>
<td>2.6</td>
</tr>
</tbody>
</table>
Finally, in the personal grid, no system administrator has access to any of the records in the database. System administrators only can access the system portion of the key for each record; the user portion is supplied by each user and is unavailable except when the user is logged in. Two separate system administrators are required to initiate a search operation, and no users (even system administrators) have access to any of the search servers when they are operating.

Since this architecture is designed to improve security, it is not surprising that its BRM value is between 1.3 and 2.6 depending on how many search servers are available for allocation. Note that this is several orders of magnitude lower than the 5.52 to 6.0 range for the prior examples of standard relational systems.

Discussion

A major security issue for health information systems is the potential for large-scale loss of data from a single unauthorized intrusion. These concerns have been reinforced by multiple, large-scale, widely reported examples of huge health data breaches, such as Anthem\(^\text{11}\) and Premera\(^\text{12}\). Such incidents have contributed to the widespread belief that central repositories of health records will not be trusted by consumers\(^\text{13}\). However, methods to objectively quantify this perceived risk have not been available.

In assessing the breach potential of database systems, it is generally understood that the risk is greater when the number of records available increases or the number of authentication steps needed to access records decreases. This observation is the basis for the BRM measure, which is designed to quantify this notion in an easily computable and understandable form. The BRM measure can also be viewed as an assessment of the “hacking reward/risk ratio” that provides a numerical representation of the number of records that will be accessible (the “reward”) for each authentication step (the “risk”). This is consistent with the observation that larger databases are more attractive targets for unauthorized access, and are more likely to be breached if there are fewer obstacles (i.e., authentication steps) needed for entry.

The results for the examples above show the usefulness of the BRM. For a standard relational database system, where a single authentication of a system administrator can provide access to the entire one million record database, the BRM value is 6. Adding an additional authentication step for data access only reduces the value to 5.7, while separating the demographic from non-demographic data results in a value of 5.52. This is consistent with the modest improvement in security provided by these alternative access arrangements.

However, when a substantially different architecture designed to improve security (the personal grid) is evaluated, the BRM measure drops dramatically to between 1.3 and 2.6 (depending on the number of search servers used). This shows clearly the security advantages of the personal grid, which separately stores and encrypts each patient’s record, thereby eliminating any path to access of the entire dataset in unencrypted form. This quantification of the security improvements for such an architecture could be very helpful to any organization considering its use.

Note that the security improvements of the personal grid result in longer search times and higher costs than an equivalent relational database. While retrieval of individual records (e.g., for clinical care) remains immediate, searching across 1 million records in a personal grid with 500 parallel processors is estimated to require about 80 seconds and cost $0.44\(^\text{10}\). For a fixed number of processors, both search times and costs will vary in direct proportion to the total number of records. However, the cost does not change with the number of processors since the overall CPU time required (which is the basis of the cost) is independent of how many search servers are allocated. Thus, personal grid search times may be reduced at no additional cost by utilizing the maximum number of parallel servers available.

In terms of the desiderata for such a measure described above, the BRM appears to meet all the criteria. It is consistent with the intuitive notion that risk is higher when more records can be exposed with fewer steps. Its quantitative nature inherently results in higher values when the risk is higher and vice versa. Since it is expressed as a logarithmic value, it can describe a very wide range of risks over many orders of magnitude. It is easy to calculate and does not depend on the specific characteristics of various system configurations. Finally, as has been shown by the examples described above, it is straightforward to interpret.

Limitations

The BRM measure for security assessment provides a specific, limited assessment of breach risk. It is not designed or intended to provide a comprehensive assessment of security risk, as it ignores several important factors that may contribute to security vulnerabilities including, but not limited to:
1. Strength of passwords/encryption. The BRM measure assumes that each authentication step functions as an effective barrier to system access and does not account for stronger or weaker passwords or encryption methods. It also does not consider the use of automatic account locking after multiple failed login attempts, or multi-factor authentication techniques. A possible variation on the BRM calculation would be to count each authentication factor separately, so two-factor authentication would be equivalent to two separate authentication steps.

2. Time factors. When data is distributed or stored inefficiently, it may take longer to retrieve. As a result, an attacker attempting to copy all the data may require an extended time period, making detection of the attack more likely. The BRM measure does not account for this possibility.

3. Other search risks. When data is distributed and must be fetched from other (outside) systems, the fetch process exposes the data to additional risks of interception or alteration. Such potential risks are not included in the BRM measure.

4. Personnel risk. Most data breaches involve some level of cooperation from insiders. The BRM measure does not assess the reliability of personnel, security training that may be provided, or the screening processes used for hiring.

5. Direct attack on unencrypted data. The BRM measure assumes that data are encrypted and cannot be accessed via an “out of channel” process that evades the normally required authentication procedures.

6. Number of users with maximum record access. In calculating the BRM, only categories of users are evaluated. The number of users having “maximum access” to records in a system is not considered. The BRM is based on the user category that provides the maximum access with the minimum number of authentication steps. Clearly, the larger the number of users in this “maximum access” category, the more likely it is that one such user’s credentials could be compromised.

7. Organizational, physical, and policy issues. The BRM measure does not include these potentially important factors.

In essence, the BRM measure only provides an estimate of the security risk of a specific data architecture and its associated access protocols. It will yield useful comparative results only if all other factors, including those described above, are held constant. Other methods should be used to assess security risk issues beyond the scope of the BRM measure.

Finally, in situations where the underlying database architecture of an application cannot be changed (e.g., when the database is a tightly integrated feature of a commercial EHR system), the BRM would only indicate the current level of risk, but would not be able to guide actions that can reduce that risk. In such cases, the BRM may be of limited value.

Conclusion

The breach risk magnitude (BRM) quantitatively measures the potential security vulnerability of databases. For a given system, it is calculated as the maximum value of the common logarithm of the number of accessible database records divided by the number of authentication steps needed to achieve such access. Estimating security risk for data organization and access arrangements and evaluation of the impact of alternatives can be facilitated using the BRM measure.

References

7. Layman L, Basili VR, Zelkowitz MV. A methodology for exposing risk in achieving emergent system properties. ACM Trans Software Eng and Methodology 2014;23(3):Article 22 (28 pages)
Understanding and Visualizing Multitasking and Task Switching Activities: A Time Motion Study to Capture Nursing Workflow

Po-Yin Yen, RN, PhD,1,2 Marjorie Kelley, RN, MS,3 Marcelo Lopetegui, MD, MS,1,4 Amber L. Rosado, RN, BSN,2,3 Elaina M. Migliore, RN, BSN,2,3 Esther M. Chipps, RN, PhD,2,3 Jacalyn Buck, RN, PhD2,3

1Department of Biomedical Informatics, 2Wexner Medical Center, 3College of Nursing, The Ohio State University, Columbus, OH, 4Clínica Alemana de Santiago, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile

Abstract
A fundamental understanding of multitasking within nursing workflow is important in today’s dynamic and complex healthcare environment. We conducted a time motion study to understand nursing workflow, specifically multitasking and task switching activities. We used TimeCaT, a comprehensive electronic time capture tool, to capture observational data. We established inter-observer reliability prior to data collection. We completed 56 hours of observation of 10 registered nurses. We found, on average, nurses had 124 communications and 208 hands-on tasks per 4-hour block of time. They multitasked (having communication and hands-on tasks simultaneously) 131 times, representing 39.48% of all times; the total multitasking duration ranges from 14.6 minutes to 109 minutes, 44.98 minutes (18.63%) on average. We also reviewed workflow visualization to uncover the multitasking events. Our study design and methods provide a practical and reliable approach to conducting and analyzing time motion studies from both quantitative and qualitative perspectives.

Introduction
Nursing or clinical workflow describes a wide range of steps or activities that health personnel execute to accomplish an activity in patient care. A fundamental understanding of multitasking within the workflow of nurses is important in today’s dynamic and complex healthcare environment. Demands for healthcare and nursing care to become a more patient-centered, efficient, effective and cost effective process, while also producing high quality patient outcomes can result in work that is fragmented due to interruptions or multitasking. Kalish and Abersold (2010), identified nurse’s work as complex and error prone. Although nurses may manage interruptions or multitasking well, it is important to gain further understanding of nursing workflow in order to develop strategies to minimize errors and maximize processes that enhance patient safety. As techniques to measure nurses’ workflow are limited, we reviewed the definition of multitasking and time motion studies being conducted commonly to model clinical workflow.

Multitasking and task switching
Multitasking requires a conscious shift in attention over a short time span between different tasks. The definition of multitasking depends on context. Lay understanding of multitasking is the performance of two or more tasks simultaneously. Healthcare researchers often define multitasking in similar ways. The related concepts of concurrent multitasking, dual task performance, parallel task performance, interleaved multitasking, and sequential multitasking describe various aspects of the cognitive processes believed to be associated with multitasking. Because these concepts overlap and intertwine, below we reviewed the definition and established the scope for our observational study.

Walter (2015) investigated on the concept of multitasking and found that multitasking has been defined in various ways; it could be concurrent multitasking (or dual task), interleaved multitasking (also called task switching), or sequential multitasking. Functional MRIs showed evidence for multitasking and task switching and the use of computer models of stimulus response activity helped our understanding of cognitive processes in the laboratory setting. Cognitive process theorists proposed that multitasking may actually be rapid task switching. Recent work in neuroimaging, eye tracking and computer simulation demonstrated that multitasking could be considered as the rapid cognitive process of task switching. However, cognitive processes examination, using fMRI, eye tracking, or computer simulations are unobservable in the real world setting.

Task switching has also been described as a rudimentary function of cognitive control requiring the flexible ability to configure and reconfigure tasks to meet shifting internal and external demands and cues. Draheim (2016) define task switching as “the ability to allocate attentive resources to several tasks sequentially and fluently reallocate
attentive resources from one task to another.” Task switching involves the performance of two or more different tasks in a rapidly integrated process and depends on several factors including complexity of task as well as familiarity of task. The daily workflow of nurses requires the performance of various skills or activities, such as communication, education, medication administration, electronic and paper documentation, as well as frequent and rapid shifts between these activities. Doing more than one thing at a time is colloquially referred to as multitasking and has been defined as “simultaneous performance of two discrete tasks.” In our study, we were unable to observe the rapid cognitive processes associated with task switching in a naturalistic setting. Therefore, we operationalized our definition of multitasking as the observable performance of two or more tasks simultaneously, for example, talking to a patient and preparing medication. Conversely, we defined task switching as alternating or changing between two separate tasks, sometimes rapidly but observably. We limit our definitions to observable behavior, as opposed to cognitive processing.

Studies measuring multitasking and task switching

Time-Motion Studies, or Time and Motion Studies, have been successfully adopted as a working method to describe and assess clinical workflow in healthcare environments. The National Library of Medicine’s controlled vocabulary thesaurus defines a Time-Motion Study (TMS) as “the observation and analysis of movements in a task with an emphasis on the amount of time required to perform the task.” TMS consists of two major components: time and motion analysis. Motion analysis represents workflow, which is a sequence of “steps” or “tasks” for a process or an event; time analysis focuses on time duration of the tasks or the entire process. Time-motion studies have been commonly used to discover nursing staff work patterns, workflow efficiency and multitasking.

Measuring multitasking or task switching enables healthcare organizations to improve efficiencies, quality and safety, workflow, and clinician job satisfaction. Various methods have been used in the study of multitasking and task switching, including qualitative studies and observational studies using paper, pencil and stopwatches, post observational interviewing, observational studies using electronic data collection devices as well as mixed-methods studies using an excel spreadsheet for data collection during the observation phase.

Although several studies have been published that have similar aims of measuring multitasking and task switching in the clinical setting, each has a slightly different design with respective strengths and limitations. Using electronic methods, Westbrook (2007) developed a Personal Data Assistance (PDA) program, called Work Observation Method By Activity Timing (WOMBAT), that allowed observers to select and record behavior from a list of categories. These categories included: direct care, indirect care, medication tasks, documentation, professional communication, ward related activities, in transit, supervision, social or non-work communications, and other tasks outside of these categories. The PDA software allowed the observer to document both multitasking (adding tasks occurring at the same time) and interruption (pausing tasks that had been interrupted) occurrences. This method provided detailed results with a high level of accuracy secondary to time stamping and duration recording functions of the software. However, it lacked the ability to record a detailed picture of each task switching or multitasking event; rather, it simply allowed the observer to denote that an event falling into one of these two categories occurred (task switching or multitasking). Lack of a universal definition of “interruption” and the inability to reliably capture these interruptions, limited the trustworthiness of the data.

Bastian (2016) conducted a time motion study focused on determining workflow (task switching) patterns in the healthcare setting, which utilized Excel spreadsheets to record observed data. This method required observers to enter an activity ID into the Excel template. All other cells populate automatically with time stamp, duration, and full task ID name. Graphical visualizations include pie and bar charts generated from Excel for the percentages in terms of time spent on tasks in each task category. Although user friendly, this data collection method failed to identify the complex details associated with task switching. Also, because of the automatic timestamp (associated in a one-to-one manner with one task), task switching and multitasking failed to be captured, thus limiting the interpretability of nursing activities from the data.

Non-electronic methods of measuring multitasking and task switching have also been used, although considerably less reliable due to manual recording processes necessitating time away from the observation. For instance, traditional stopwatch and paper methods have been used to record multitasking and task switching of clinicians. In a time motion study conducted by Edwards (2009), observers shadowed clinicians for two to four hour periods using real time stopwatches and a paper/pencil observational tool system. Another study focused on the changes in communication tasks rather than patient care functions, in which written documentation and audio recordings were used to record
communication events. The audio recordings were used to accurately monitor interruptions and were later analyzed and transcribed. Although effective in recording communication, multitasking, and task switching activities, observers were responsible for both timing the duration of events and providing detailed written accounts. Observation accuracy was limited because of time necessary for paper-pencil documentation. Additionally, non-verbal communications failed to be captured. Lastly, Berg (2013) utilized interviews along with handwritten documentation of observations to obtain clinicians’ perspectives on their experiences with workflow interruptions. In this study, the observers - experienced nurses - documented six categories (type of activity, duration of the activity, location, persons, interruptions, whether the clinician was the recipient cause of the interruption, and continuation of activity after interruption), which were recorded in written form on a semi-structured template. While audio recordings and post-observation interviews may be useful adjunct methods to an electronic time-motion methods, handwritten documentation methods lack reliable because of the time required of the observer to record data rather than fully observing clinicians behavior.

In summary, studies using a time motion method to study clinical workflow were inconsistency in their approaches. Zheng (2011) proposed Suggested Time And Motion Procedures (STAMP) to standardize research for time motion studies. Following STAMP, we designed a time motion study to understand and visualize nursing workflow on multitasking and task switching activities. We used TimeCaT, a comprehensive electronic time capture tool to capture data for our time-motion study. TimeCaT is a web-based application, optimized for tablets and iPads developed by the Department of Biomedical Informatics at The Ohio State University. It is a fully customizable, web-based, time capture tool designed for time motion studies. TimeCaT can be used to evaluate multitasking, task switching, and inter-observer reliability, and can be used on any internet capable device. TimeCaT has also been optimized to support data capture using touch enabled devices. We loaded selected nursing activities on TimeCaT (Figure 1), trained student observers, and validated the data capture process using the inter-observer reliability module included in TimeCaT.

Our study was a pilot project for a larger study aiming to understand top of license nursing practice. The purpose of this pilot project was to establish the rigor of the time motion method. With the preliminary data, we were able to describe nursing workflow in three activity dimensions: communication, task, and location. Communication includes with whom nurses are interacting; hands-on tasks include tasks requiring nurses to physically touch to perform a task (i.e. start an IV); and location includes where nursing activities take place. We explored these three activity dimensions, across the continuum of time to understand the phenomena of multitasking and task switching in nursing practice.

Methods
Setting and Sample
After obtaining IRB approval, the study was conducted on a medical-surgical unit at The Ohio State University Wexner Medical Center (OSUWMC). The observations occurred in the general patient care areas including the nursing station, hallway, medication room, patient room, and supply areas. This was a pilot study in which we used a convenience sample of registered nurses. The nurses met the following inclusion criteria: (1) full-time staff Registered Nurses (RNs) working at OSUWMC with more than two years of acute care nursing work experience and (2) greater than or equal six months of work experience on the study unit. Informed consent was obtained for this observational study at the time of recruitment.

Observed nursing activities
We observed all nursing activities including anticipated nursing activities and activities which may be classified as non-value added. Anticipated activities include but are not limited to; hand-off (shift reporting), direct patient care
Inter-observer reliability assessment (IORA) participates as well as other unit personnel. This also helped introduce the study and study personnel to the nurse observations. All trial observations occurred in the study site which allowed student observers to be familiar with the inter-observer reliability assessment to ensure study fidelity and data validity prior to beginning actual research. Based on the Needleman-Wunsch sequence alignment algorithm from bioinformatics, the proposed method returns a normalized score that represents an overall sequence agreement. Table 2 presents three IORA training results from two observers. Their performance was improved over time. The IORA provided useful information for observers’

Table 1. Selected activities definition table

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Definition</th>
<th>Start-time</th>
<th>End-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comm.</td>
<td>Patient Education</td>
<td>Planned patient education; direct communication with patient</td>
<td>RN starts talking about education content</td>
<td>RN stops talking about education content or switches to another topic</td>
</tr>
<tr>
<td>Comm.</td>
<td>Patient Non-education</td>
<td>Direct communication with patient not concerning planned education; any communication other than planned education.</td>
<td>RN or patient starts talking</td>
<td>Both RN and patient stop talking</td>
</tr>
<tr>
<td>Comm.</td>
<td>M.D.</td>
<td>Direct communication (verbal and non-verbal) between nurse and physicians</td>
<td>RN or MD starts talking</td>
<td>Both RN and MD stop talking</td>
</tr>
<tr>
<td>Comm.</td>
<td>Call In</td>
<td>RN answers a call on her OSUMC mobile phone.</td>
<td>RN says hello; starts talking to person calling</td>
<td>RN hangs up the phone</td>
</tr>
<tr>
<td>Comm.</td>
<td>Call Out</td>
<td>RN calls someone on her OSUMC mobile phone.</td>
<td>RN hits green call out button on phone</td>
<td>RN hangs up the phone</td>
</tr>
<tr>
<td>Hands-on Task</td>
<td>Direct - Assessment</td>
<td>RN performs direct physical assessment</td>
<td>Places hands and/or stethoscope on patient</td>
<td>Removes hands and/or stethoscope</td>
</tr>
<tr>
<td>Hands-on Task</td>
<td>Direct - Procedure</td>
<td>RN performs treatment or procedure that cannot be delegated (top of license tasks)</td>
<td>RN places hands on pt. or supplies to begin procedure</td>
<td>RN removes hands or switches to another direct task</td>
</tr>
<tr>
<td>Hands-on Task</td>
<td>Direct - Procedure - delegable</td>
<td>Direct patient care that may be delegated to a PCA or other unlicensed persons</td>
<td>RN places hands on pt. or supplies to begin procedure</td>
<td>RN removes hands or switches to another direct task</td>
</tr>
<tr>
<td>Hands-on Task</td>
<td>Transportation-prep</td>
<td>Get patient prepared to travel for procedure</td>
<td>Initiate travel preparation, gathering supplies for transport only</td>
<td>patient begins to transport (no longer preparing for transport)</td>
</tr>
<tr>
<td>Hands-on Task</td>
<td>EHR-Charting</td>
<td>Recording, entering or updating data in the EHR</td>
<td>RN logs on to EHR and starts typing in EHR</td>
<td>RN stops typing or logs off computer</td>
</tr>
<tr>
<td>Location</td>
<td>Travel/Walking</td>
<td>Time spent walking/traveling in between destinations</td>
<td>Leaves destination and begins traveling</td>
<td>Reaches destination and stops traveling</td>
</tr>
<tr>
<td>Location</td>
<td>Own Patient Room</td>
<td>In nurses’ assigned patient rooms</td>
<td>Enters assigned patient room and crosses doorway</td>
<td>Leaves assigned patient room and crosses the doorway</td>
</tr>
<tr>
<td>Location</td>
<td>Hallway</td>
<td>In hallway as a destination</td>
<td>Stops walking in hallway</td>
<td>Continues to walk</td>
</tr>
</tbody>
</table>

Student observers

In order to obtain quality observational data, we recruited three nursing student as our observers, including one nursing PhD student and two undergraduate senior nursing students. With their background and clinical experience, they were familiar with nursing workflow and were able to recognize and distinguish various nursing activities. All observers understood the study purpose, and were required to participate in research meetings in which of the nursing activities definitions were established. Observers were trained with trial observations for at least 12 hours, and three rounds of inter-observer reliability assessment to ensure study fidelity and data validity prior to beginning actual research observations. All trial observations occurred in the study site which allowed student observers to be familiar with the environment as well as the on-site observation. This also helped introduce the study and study personnel to the nurse participates as well as other unit personnel.

Inter-observer reliability assessment (IORA)

The establishment of IORA protocols and guidelines is the priority in validating a time motion study. We used the IORA provided in TimeCaT,35 a robust, comprehensive method for inter-observer reliability assessment.37 It considers four types of agreements on the workflow observation: 1) proportion-kappa (P-K): evaluates the naming agreement on virtually created one-second activity, based on the Kappa statistic. P-K provides a global assessment of the agreement over time; 2) naming-kappa (N-K): a systematic pairing approach based on time-overlap, and provides a Kappa statistic representing the agreement on activity naming; 3) duration-concordance correlation coefficient (DCCC): agreement on the duration of an activity. It is the concordance correlation coefficient and intends to provide a means to evaluate the correlation concerning activity duration; and 4) sequence-Needleman-Wunsch (S-NW): agreement of the sequence of activities. It is an assessment focusing specifically on the order of activities recorded. Based on the Needleman-Wunsch sequence alignment algorithm from bioinformatics, the proposed method returns a normalized score that represents an overall sequence agreement. Table 2 presents three IORA training results from two observers. Their performance was improved over time. The IORA provided useful information for observers’
training and a meaningful quantitative IORA report. In addition, we visualized clinical workflow by their sequence and duration, as presented in Figure 2. Each colored band represents an activity; the width of the band represents the duration of the activity. For example, EHR charting and EHR review are two activities which occurred more frequently and longer in duration than other activities. We were also able to compare the activities observed by Observer A against to the activities observed by Observer B. The side-by-side workflow visualization (Figure 2) in TimeCaT allowed us to offer immediate feedback to our observers after the training sessions.

Table 2. Sample IORA training results from two observers

<table>
<thead>
<tr>
<th></th>
<th>Communication</th>
<th>Hands-On Task</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-K</td>
<td>N-K</td>
<td>D-CCC</td>
</tr>
<tr>
<td>IORA1</td>
<td>83.41%</td>
<td>73.97%</td>
<td>0.257P</td>
</tr>
<tr>
<td></td>
<td>0.71K</td>
<td>0.65K</td>
<td>0.59K</td>
</tr>
<tr>
<td></td>
<td>85.16%</td>
<td>66.73%</td>
<td>0.906p</td>
</tr>
<tr>
<td></td>
<td>0.78K</td>
<td>0.59K</td>
<td>0.966p</td>
</tr>
<tr>
<td></td>
<td>95.02%</td>
<td>77.41%</td>
<td>0.997p</td>
</tr>
<tr>
<td></td>
<td>0.94K</td>
<td>0.69K</td>
<td>0.997p</td>
</tr>
<tr>
<td></td>
<td>66.73%</td>
<td>0.59K</td>
<td>0.966p</td>
</tr>
<tr>
<td></td>
<td>0.78K</td>
<td>0.59K</td>
<td>0.966p</td>
</tr>
<tr>
<td></td>
<td>77.41%</td>
<td>0.69K</td>
<td>0.997p</td>
</tr>
<tr>
<td></td>
<td>0.94K</td>
<td>0.69K</td>
<td>0.997p</td>
</tr>
<tr>
<td>IORA2</td>
<td>85.51%</td>
<td>61.31%</td>
<td>0.98p</td>
</tr>
<tr>
<td></td>
<td>0.7K</td>
<td>0.46K</td>
<td>0.897p</td>
</tr>
<tr>
<td></td>
<td>85.14%</td>
<td>56.8%</td>
<td>0.897p</td>
</tr>
<tr>
<td></td>
<td>0.73K</td>
<td>0.46K</td>
<td>0.897p</td>
</tr>
<tr>
<td></td>
<td>87.3%</td>
<td>60.14%</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>0.85K</td>
<td>0.49K</td>
<td>0.987</td>
</tr>
<tr>
<td>IORA3</td>
<td>93.34%</td>
<td>71.03%</td>
<td>0.954p</td>
</tr>
<tr>
<td></td>
<td>0.8K</td>
<td>0.59K</td>
<td>0.992p</td>
</tr>
<tr>
<td></td>
<td>88.87%</td>
<td>73.21%</td>
<td>0.992p</td>
</tr>
<tr>
<td></td>
<td>0.79K</td>
<td>0.67K</td>
<td>0.992p</td>
</tr>
<tr>
<td></td>
<td>97.46%</td>
<td>84.85%</td>
<td>0.998p</td>
</tr>
<tr>
<td></td>
<td>0.97K</td>
<td>0.79K</td>
<td>0.998p</td>
</tr>
</tbody>
</table>

P-K= proportion-kappa; N-K= naming-kappa; D-CCC= duration-concordance correlation coefficient; S-NW= Sequence-Needleman-Wunsch

Procedure

Student observers obtained informed consent from the observed nurse as well as permission from patients for observing their care. Each observation lasted for four hours over preset time periods: 7am to 11am, 11am to 3pm, or 3pm to 7pm, depending on availability arrangement. No identifiable information or health records were collected. The study was approved by the Institutional Review Board (IRB) at The Ohio State University.

Results

We completed a total of 56 hours (fourteen valid 4-hour observations) with 10 registered nurses, including eight observations from 11am to 3pm, five observations from 3pm to 7pm, and one observation from 7am to 11am. Among the 14 observations, three were on Monday, one on Tuesday, one on Wednesday, five on Thursday, and three on Friday.

Multi-tasking

In a 4-hour observation time block, nurses had 124 communications and 208 hands-on tasks on average. We considered that nurses were multitasking when they were engaged in communication and hands-on tasks simultaneously. We found that nurses were multitasking 131 times on average, which is 39.48% of all times; the total multitask duration ranges from 14.6 minutes to 109 minutes, 44.98 minutes (18.63%) on average. Multitasking duration seemed higher from 7am to 11am, but because we had only one observation during that time frame, we were not able to be conclusive.

Table 3 shows more details for each observation and the proportion of multitasking in frequency and duration (second). For example, Observation 1 occurred from 11am to 3pm, a 4-hour observation. During the observation, the nurse communicated 132 times with others and performed 171 hands-on tasks. The nurse multitasked (having communication and hands-on task simultaneously) 105 times, which is about 34.65% of all times. The duration of multitasks is 37.97 minutes, 15.76% of the total duration.
We also examined workflow visualization (Figure 3) to explore the location of nursing activities occurred. We presented the workflow in three activity dimensions: communication, hands-on task, and location. We visualized these three activity dimensions across the continuum of time with colored bands representing different activities; the width of the band represents the duration of an activity. The activities on the three activity dimensions, and at the same horizontal position, represent multitasking. The workflow visualization can be interpreted as who the nurse was talking to, while doing what, and at what location. We found that nurses usually charted at the nursing station or on the hallway. They sometimes charted while they were traveling/walking to another destination. When they were reviewing or charting on the electronic health record, they were often talking to their patients or to other professions. Nurses usually called out when reviewing or charting, but could receive phone calls at anytime and anywhere. Task switching often occurred between related activities, such as reviewing and charting electronic health records, switching between direct patient care activities (e.g. procedure, vital sign, medication administration). Another example of related activities was calling out and electronic health record review. Calling out required looking up pager or phone numbers using the electronic health record, so these two activities were often closely linked.

**Discussion**

**Self-created multitasking vs. unexpected interruption**

During this pilot study, we observed that multitasking and task switching could be self-directed, such as consulting with other professions while reviewing patient records, talking to patients for assessment and documenting results simultaneously, charting while walking, and assessing patients while administering medications. Self-directed multitasking and task switching were considered being more efficient at work and are likely contributing to their job satisfaction. Conversely, we also observed unexpected multitasking or interruptions, such as receiving phone calls anytime and anywhere, or mini-conversations with others while preparing medications. Some interruptions are important as they could be reminders, alerts, or another higher priority task. However, interruptions also increase cognitive load, decrease the speed of information processing in human prefrontal cortex, and result in errors. While multitasking is necessary and an important skill in healthcare systems, it is important to assess and recognize its impact on patient care. For example, allowing patients direct contact with nurses, through the use of pagers, could increase patient satisfaction, but it may also increase unexpected interruptions. Such interruptions could potentially contribute to safety issues, particularly if nurses are performing critical procedures or preparing complex medication regimens.
Currently, no model exists to evaluate multitasking. Multitasking could be self-directed or not, and could increase or decrease efficiency. To help explicate these workflow nuances, we used a workflow visualization method to understand the phenomena. We defined multitasking and task switching and designed the time motion study to capture observable nursing activities from three activity dimensions: communication, hands-on tasks, and location over the continuum of time. Using TimeCaT, we were able to analyze nursing workflow from the timing and frequency of the three activity dimensions and investigate the context and rationale via workflow visualization. Our study design and method provides a practical and reliable approach to conducting and analyzing time motion studies from both quantitative and qualitative perspectives.

Figure 3. A Snapshot of nursing workflow visualization

Visualizing nursing workflow: communications, hands-on tasks, and locations of nursing activities
Study limitations

Observing nurses in one unit in one hospital within one academic health system limits generalizability of our study. Data collection occurred over a two-month period which may introduce bias due to seasonal variations in hospital admissions. Participation bias was also a limitation, as nurses volunteered to participate in the study. Also, observations could be limited due to that humans may not be able to accurately record every action that is occurring, especially in a high stress clinical situation. We may have missed collection of data, particularly in rapidly changing hands-on tasks, communications or locations/movements of nurses. To minimize observational bias, all our observers were required to establish IORA before their observation data could be considered as valid thus also insuring study fidelity. We could not guarantee 100% accuracy, but we minimized the inconsistency between observers through rigorous training and clearly defining definitions of communications, hands-on tasks, and locations.

Future direction

In our study, we only recorded nursing activities that were observable. However, nurses often multitask and task switch in their thinking – critical thinking – which is unobservable. For example, nurses may juggle various needs of patients, family, and co-workers. Future research should investigate the impact of multitasking by assessing perceived workload at the end of observation to strengthen study findings. In addition, working within the complex healthcare environment of today, nurses must provide care that is efficient and effective. There is a growing body of literature that describes the workflow of nurses in an acute care environment. It has been found that nurses spend a considerable amount of non-value added time on activities that could potentially be delegated to other team members who could accomplish the care safely and with greater cost effectiveness. Inefficiencies in organizational systems also contribute to non-value added time. To date, there has been no empirical work that has aligned nursing practices and workflow or examined these practices relative to the notion of “top of licensure”. Top of license nursing practice addresses how nurses spend their time across the care continuum and examines “non-valued-added” work which could be executed safely by other healthcare personnel. As this is a pilot project under a larger nursing practice study, our ongoing and future research will explore top of license nursing activities and provide a strategic solution for providing higher quality and more efficient care.

Conclusion

We conducted a time motion study to capture communications, hands-on tasks, and locations of nursing activities. We also described our methodology in detail, and demonstrated a practical and reliable approach for other researchers. With the observational data, we were able to identify the amount of multitasking being carried out by nurses in their daily work, and provide a vivid pictures of nurses’ activities via a workflow visualization. Even though multitasking sometimes cannot be avoided, it has been noted that “Multitasking, a media-driven bias toward dramatic scenarios, and an emphasis on meeting institutional goals in the form of documentation have led to a culture of action-based practice, which interferes with nurses’ ability to simply be with patients. In order for nurses to be fully present with their patients, the cultural norm of multitasking and the emphasis on doing must be reexamined within the context of patient care.” We plan to continue the time motion study with a larger sample of nurses to observe their workflows and activities, with the hope of providing a quantitative observational study with statistically rigorous evidences to describe the work of nurses. Future research also includes applying the approach and methodology to investigate the workflow change before and after the implementation of a new practice model or a new technology.

References


38. Forsberg HH, Muntlin Athlin A, von Thiele Schwarz U. Nurses' perceptions of multitasking in the emergency department: effective, fun and unproblematic (at least for me) - a qualitative study. Int Emerg Nurs. 2015;23(2):59-64.
Moderating the Influence of Current Intention to Improve Suicide Risk Prediction

Nawal A. Zaher, MSc¹, Christopher D. Buckingham, PhD²
¹AASTMT, Cairo, Egypt; ²Aston University, Birmingham, United Kingdom

Abstract

When assessors evaluate a person's risk of completing suicide, the person's expressed current intention is one of the most influential factors. However, if people say they have no intention, this may not be true for a number of reasons. This paper explores the reliability of negative intention in data provided by mental-health services using the GRiST decision support system in England. It identifies features within a risk assessment record that can classify a negative statement regarding current intention of suicide as being reliable or unreliable. The algorithm is tested on previously conducted assessments, where outcomes found in later assessments do or do not match the initially stated intention. Test results show significant separation between the two classes. It means suicide predictions could be made more accurate by modifying the assessment process and associated risk judgement in accordance with a better understanding of the person's true intention.

Keywords: Suicide Risk, Suicide Intention, Clinical Risk Judgment, GRiST, Decision Support System

Introduction

Intention is a mental state that represents a commitment to carrying out an action or actions in the future. It is one of the most important components for assessing the risk of a suicide attempt by mental health practitioners but it is also difficult to measure. It relies on people being assessed giving accurate self reports and there are many reasons why this may not happen. The problem is not so much with those who indicate a degree of positive intention because this alerts the assessor to the need for exploring intention further. Intentions can be evaluated as representations of possible actions or plans set to achieve a goal; realistic and established plans with steps taken towards the goal indicate increased underlying intention.

While it is reasonable to claim that the presence of current intention of suicide is an indication of risk, the absence of intention does not necessarily dictate the opposite. Patients may, knowingly or unknowingly, hide their intentions, which raises reasonable doubt regarding their statement of intention.

The difficulty with people who say they have no intention is that validating their statement depends on an absence of evidence and this is difficult for assessors to validate. In effect, it stops further investigation into the current intention, diverts attention elsewhere, and can mean important information is lost. The overall risk judgement is compromised as well as a comprehensive understanding of the underlying symptoms and causes. The assessor’s evaluation plan and course of action are affected, which results in a mismatch between interventions and outcomes. It happens with a “no” answer because of its categorical nature, as opposed to a “yes” answer, which can be further investigated to attenuate its influence on risk judgements.

This study uses data from a web-based decision support system (DSS) called the Galatean Risk and Safety Tool, GRiST to determine the reliability of people’s expressed absence of suicide intention. The goal is to identify cues that would enable assessors to change how the absence of current intention is conceived. If it is possible to detect unreliable absence of intention, the assessment process can be changed to explore it further or, at the very least, the influence of the stated absence of intention on the assessor’s judgement of suicide risk can be reduced.

The next section provides an overview of the nature of mental health risk data and how it is collected by GRiST. The paper then introduces an algorithm to classify people’s negative intentions of suicide as being reliable or unreliable. The results are presented and followed by a discussion about how the algorithm could be incorporated in risk evaluation tools to improve their accuracy.

Background

Questions about patients’ intentions direct assessors and DSSs towards the cues that are most relevant to the risk being assessed; some or all of these could be masked if people have stated an absence of intention. It leads to two major problems. The first is missing data, where relevant data is not collected because the assessor thinks it is not needed. The second concerns outcomes because people wrongly assessed as low risk may not receive required

1274
interventions and can go on to make a suicide attempt. Assessments are made more difficult by the volatile nature of
stated intentions, which is compounded by factors such as impulsiveness and deception or denial. In short, we lack
insight into the intention of others\textsuperscript{15}. The GRiST DSS attempts to penetrate the gloom by collecting a large amount
of contextual and historical data, in addition to the immediate risk history and behaviour, so that a holistic picture of
a person can be drawn. This is the data set used for modeling and evaluating current intention.

1- GRiST Mental Health Data

GRiST is a tool for helping practitioners assess and manage multiple risks associated with mental-health problems,
including suicide, self-harm, harm to others, self-neglect and vulnerability. The tool represents expert consensus,
elicted by preliminary interviews conducted with mental health practitioners\textsuperscript{16} and refined through feedback from
using the tool in practice\textsuperscript{13,17}. The level of suicide risk is measured in GRiST using membership grades (MGs). They
are generated from the person’s assessment data into a common value between zero and one that represents each cue’s
risk input\textsuperscript{12}. The relationship between input MGs and associated risk judgements given by assessors is learned across
all assessments in the GRiST database to provide a model for predicting the risk judgement for a new assessment\textsuperscript{12,13}.

2- Current Intention of Suicide

The clinical risk judgment distributions for GRiST assessments that have “yes” and “no” answers (Figure 1) show
that the mean risk for the “yes” group is 0.5201 and for the “no” group is 0.1764. This is a significant clinical
separation, with people being treated differently when having more than a three-point greater risk level, where 0 is
minimum risk and 1 is maximum risk.

It could be argued that intention is more accurately predicted by attitude and behaviour\textsuperscript{3,18}, rather than by a simple
YES/NO question but their ability to interpret the reliability of someone’s lack of current intention has not been
modeled. For GRiST, current intention is not binary for a “yes” answer because it opens up a number of additional
questions to assess the degree of intention. In contrast, a “no” answer blocks off any further direct exploration of
current intention and may also cause other factors to be masked because of a perceived reduction in risk. An indirect
approach to evaluating the “no” answer is required and this paper pursues it by investigating whether any of the
multiple GRiST cues can differentiate between the reliability of assessments’ claimed lack of intention.

Proposed Algorithm

Although GRiST does not directly collect intervention and outcome information following an assessment, it does
contain relevant data for those patients who have more than one assessment. An important piece of information
recorded for all patients who have a history of suicide attempts is the date of the most recent attempt. If this has

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Distribution of clinical risk judgment for patients with current intention of “yes” (the top figure) and “no”
(the bottom figure).}
\end{figure}
changed between assessments then it means the person has carried out a new suicide attempt and is therefore an outcome we can use in our analysis.

This paper analyses all those assessments that had no current intention and are followed by a subsequent (repeat) assessment. The idea is that people with a new risk episode will have a less reliable absence of current intention than those who did not carry out an attempt between assessments. If the two groups of patients can be distinguished, it should be possible to determine a reliability measure for the negative answer and a principled way of adjusting the assessment process accordingly.

1- Defining the Classes

The parent population for the study sample is 71,024 assessments, consisting of 27,947 patients, of which 12,595 have repeated assessments. Of these, only 6,502 assessments have a current intention answer of “no” in an assessment which also has a subsequent one (i.e. it must not be the most recent assessment). The reliability of the answer is crudely evaluated by inspecting the subsequent assessment to find out whether this patient has had a repeat episode since the assessment, indicated by a new date for the most recent attempt. If so, the answer to current intention is marked as unreliable. This produces two classes: those people with a reliable negative intention, Class Rno, and those with an unreliable answer of “no”, Class Uno. The size of Class Rno is 4,458 assessments, and the size of Class Uno is 2,044 assessments.

2- Feature Selection

Having established the two classes, the task is to learn the defining features that distinguish them. The Maximum Likelihood Estimate (MLE)\textsuperscript{19} of the mean and variance of each feature in each class is calculated assuming a Gaussian distribution. MLE provides an interpretable parametric form and has a significantly low computational cost compared to other parametric techniques like Bayesian estimation\textsuperscript{20}. Then the normalized Euclidean distance\textsuperscript{21} between the two classes is calculated for each feature to produce the distance vector shown by Equation 1

\[
d = \sqrt{\frac{\left(\bar{\mu}_{\text{Rno}} - \bar{\mu}_{\text{Uno}}\right)^2}{\bar{\sigma}_{\text{Rno}}^2 + \bar{\sigma}_{\text{Uno}}^2}}
\]

where \(\bar{\mu}_{\text{Rno}}\) and \(\bar{\mu}_{\text{Uno}}\) are the average assessments of class Rno and class Uno respectively and \(\bar{\sigma}_{\text{Rno}}^2\) and \(\bar{\sigma}_{\text{Uno}}^2\) are the variances of Rno and Uno. Each element of the distance vector represents the distance between Rno and Uno for a particular cue.

The top 20 cues arranged in descending order of the distance measure, d (Table 1), are appealing candidates for the feature vector. Although these produce the largest class separation, some will be missing from assessments if the assessors do not ask about them due to absence of intention. The table gives the number of occurrences of each of the cues in the dataset and the percentage of assessments where they are missing. The cues with the least amount of missing data and the most influence are the first seven cues. If only these are used, the classifier will have a larger sample and be more reliable because only assessments with data for all the cues in the vector are used for the estimation of model parameters and for testing. These cues all happen to be related to previous history of suicide, which now becomes a selection criterion for the unreliability measure. It means a different measure will need to be calculated for other subsets of the population that do not have a history of suicide.

It is encouraging, but not surprising, that the most influential cues are also the most commonly collected. GRiST was built on the pooled expertise of mental-health practitioners\textsuperscript{16} and one would expect them to ensure they identify and collect the most important data for assessments. The literature also confirms the importance of cues related to past history for predicting suicide risk\textsuperscript{4,13,22,23}.

<table>
<thead>
<tr>
<th>No.</th>
<th>Cue</th>
<th>Distance</th>
<th>Occurrences</th>
<th>Missing (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date of most recent suicide attempt</td>
<td>10.16</td>
<td>6276</td>
<td>3.50%</td>
</tr>
<tr>
<td>2</td>
<td>How much did the person want to succeed</td>
<td>2.70</td>
<td>5873</td>
<td>9.70%</td>
</tr>
<tr>
<td>3</td>
<td>Chance of discovery after suicide attempts</td>
<td>2.60</td>
<td>5824</td>
<td>10.4%</td>
</tr>
</tbody>
</table>

Table 1. The top 20 cues listed in order of how well they separate the reliability classes, given by their distance measures. The number of occurrences of the cue is within the sample of 6,502 assessments, which is also given as a percentage of assessments with the cue missing.
<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>2.59</th>
<th>5713</th>
<th>12.3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Regret about trying to commit suicide</td>
<td>2.55</td>
<td>6130</td>
<td>5.70%</td>
</tr>
<tr>
<td>6</td>
<td>Insight into lethality of previous suicide attempts</td>
<td>2.41</td>
<td>6015</td>
<td>7.50%</td>
</tr>
<tr>
<td>7</td>
<td>Suicide attempts escalating in frequency</td>
<td>1.78</td>
<td>5443</td>
<td>16.3%</td>
</tr>
<tr>
<td>8</td>
<td>How many suicide attempts</td>
<td>1.14</td>
<td>3965</td>
<td>39.0%</td>
</tr>
<tr>
<td>9</td>
<td>Likelihood of acting on delusions</td>
<td>1.08</td>
<td>3230</td>
<td>50.3%</td>
</tr>
<tr>
<td>10</td>
<td>Potential triggers match previous triggers</td>
<td>0.98</td>
<td>2314</td>
<td>64.4%</td>
</tr>
<tr>
<td>11</td>
<td>Stage of depression</td>
<td>0.95</td>
<td>1818</td>
<td>72.0%</td>
</tr>
<tr>
<td>12</td>
<td>Number of dependents sharing accommodation</td>
<td>0.93</td>
<td>6308</td>
<td>3.00%</td>
</tr>
<tr>
<td>13</td>
<td>Mania/hypomania</td>
<td>0.89</td>
<td>3390</td>
<td>47.8%</td>
</tr>
<tr>
<td>14</td>
<td>Life not worth living</td>
<td>0.84</td>
<td>3252</td>
<td>50.0%</td>
</tr>
<tr>
<td>15</td>
<td>Capacity to cope with major life stresses</td>
<td>0.83</td>
<td>3028</td>
<td>53.4%</td>
</tr>
<tr>
<td>16</td>
<td>Habitable accommodation</td>
<td>0.81</td>
<td>2046</td>
<td>68.5%</td>
</tr>
<tr>
<td>17</td>
<td>General motivation in life</td>
<td>0.79</td>
<td>1013</td>
<td>84.4%</td>
</tr>
<tr>
<td>18</td>
<td>Plans for the future</td>
<td>0.74</td>
<td>3287</td>
<td>49.4%</td>
</tr>
<tr>
<td>19</td>
<td>Potential triggers of suicide</td>
<td>0.73</td>
<td>2859</td>
<td>56.0%</td>
</tr>
<tr>
<td>20</td>
<td>Impulsiveness</td>
<td>0.72</td>
<td>2981</td>
<td>54.1%</td>
</tr>
</tbody>
</table>

### 3- Classifier Model

There are many approaches to binary classification, the most prominent of which is logistic regression. Although it is powerful and appropriate, we decided against it for our approach because the heuristic method for choosing features was not designed to produce regression vectors. It could have been achieved using exhaustive search but it would have been limited to samples with all features present in the vector and over fitting may have occurred due to selecting features based solely on classification results.

Instead, features are independently chosen using a fully parametric probabilistic model to generate the conditional density functions on the assumption that each feature has a Gaussian distribution. Features do not all need to be in the same patient vector, which means selection can include any assessments where they individually occur, making the sample size for selecting each one larger than would be the case for regression.

The classes are represented by two Gaussian distributions, with means and variances calculated using MLE for the top seven cues shown above (Table 1). The classification decisions are based on a minimum error rate threshold in each dimension. An example for the classifier model for one dimension: the most recent suicide attempt date, which is the top cue, is shown below (Figure 2). In each dimension the classifier needs a threshold to separate the two classes. The optimum threshold \( \rho \) that minimizes the probability of error is the MG value at the point of intersection of the two Gaussian distributions representing the two classes (Figure 1). Equating the two sides and substituting for \( \rho \):

\[
\frac{1}{\sqrt{2\pi\sigma_{Dno}^2}} e^{-\frac{(\rho-\mu_{Dno})^2}{2\sigma_{Dno}^2}} = \frac{1}{\sqrt{2\pi\sigma_{Kno}^2}} e^{-\frac{(\rho-\mu_{Kno})^2}{2\sigma_{Kno}^2}}
\]

Simplifying the expression yields Equation 3 with the coefficients \( a_2, a_4, a_6 \) given by Equations 4, 5, and 6 respectively. The threshold \( \rho \) that minimizes the probability of error in each dimension is calculated for each dimension individually by solving Equation 3.

\[
a_2 \rho^2 + a_4 \rho + a_6 = 0 \tag{3}
\]
Figure 2. Classifier model for Cue 1, “date of most recent suicide attempt”, showing the Gaussian distribution for each class, Uno and Rno (unreliable and reliable no intention respectively), and the decision threshold for this dimension, based on the cue’s membership grade, MG.

Figure 3. Distribution of MG values for cue 1, “most recent suicide attempt”, RHS: MG values of cue 1 for class Uno, LHS: MG values of cue 1 for class Rno.
The idea is to find the input MG for each of the seven cues that split the classes most accurately. Each cue will independently indicate the most likely class by whether the assessment MG for the cue is above or below the threshold value.

The natural distribution of data shown (Figure 3) supports the Gaussian assumption adopted. Nevertheless, considering more skewed models may be of interest, but will add some complexity to parameters and threshold computations.

4- Decision Fusion

The last stage of the classification process is combining the decisions from all 7 cues (dimensions) into one decision as to whether an assessment belongs to Class Rno or Uno. This can be done by voting over the decisions of each cue or summing their probabilities that the assessment is in one class or the other. This paper combines the two methods because a majority vote among the hard decisions taken by each dimension does not take into account the accuracy of each dimension.

First a hard decision is made for each cue by comparing the value of the cue’s MG to the threshold. The cue’s vote is then weighted based on how well it separates the classes. This is determined by calculating its positive predictive value (PPV) using Equation 7

\[ ppv = \frac{tpr}{tpr + fpr} \quad (7) \]

where TPR is the true positive rate and FPR is the false positive rate. The PPV shows how powerful each dimension (cue) is individually, with a higher PPV signifying more separation of the two classes in that dimension.

The PPV value for each cue is used to weight the overall decision obtained from its corresponding dimension as in Equation 8

\[ v = \sum_{i=1}^{7} ppv_i d_i \quad (8) \]

where \( ppv_i \) is the positive predictive power of dimension \( i \) and \( d_i \) is a Boolean variable that represents the decision from each cue, which is given a value of ‘1’ for class Uno and a value of ‘0’ for class Rno.

The output of the fusion process, \( v \), is compared to a threshold \( v_{th} \) such that if \( v > v_{th} \), then the decision is Rno, otherwise the decision is Uno. To choose \( v_{th} \), the Receiver Operating Characteristics (ROC) curve is plotted using different values for \( v_{th} \) so that the point with the maximum accuracy is chosen. The figure below (Figure 4) shows the ROC curve of the best performing model and its distance from the line of no discrimination.

![ROC curve](image)

**Figure 4.** ROC curve showing TPR against FPR for different decision thresholds.
Results
After filtering out assessments with incomplete vectors (i.e. where one or more of the top seven cues in Table 1 are missing), the size of Class Rno is 1,154 and Class Uno is 1,903. The model was tested by combining the two classes into one group and then applying 10-fold cross validation, where the group is split into 10 equal-sized sets of randomly selected assessments. Each set is used once for testing with the remaining 9 sets used for training to give 10 sets of test results that were pooled.

1- Accuracy
The average accuracy of the 10 splits is 0.847 while the percentage values of TPR, FPR, false negative rate (FNR) and true negative rate (TNR) at different setups are shown below (Table 2). The point with the highest accuracy is not necessarily the best operating point. It depends on which of the cells needs to be maximized or minimized. For the sake of risk assessment, the TPR is probably the most important but the FPR is also significant because false alarms would erroneously impact on the assessment process as well as exaggerating the risk and potentially triggering unnecessary interventions.

Table 2. Classifier statistics showing average performance, best performing split and worst performing split for Class Rno. TPR is the true positive rate, or hits, FPR is the false positive rate or false alarms, TNR is the true negative rate, or hits for Class Uno, and FNR is the false negative rate or misses.

<table>
<thead>
<tr>
<th>Performance</th>
<th>Accuracy</th>
<th>TPR</th>
<th>FPR</th>
<th>TNR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>84.07%</td>
<td>72.55%</td>
<td>4.41%</td>
<td>95.59%</td>
<td>27.45%</td>
</tr>
<tr>
<td>Best</td>
<td>89.98%</td>
<td>72.62%</td>
<td>2.02%</td>
<td>97.98%</td>
<td>27.38%</td>
</tr>
<tr>
<td>Worst</td>
<td>79.51%</td>
<td>60.01%</td>
<td>11.24%</td>
<td>88.76%</td>
<td>39.99%</td>
</tr>
</tbody>
</table>

2- Chi-square test
To verify the significance of the results a Chi-square test was performed to check if the classification decision (Rno or Uno) is dependent on whether the person subsequently repeated a suicide attempt or not. The sample sizes on which the test was performed and the results of the classification at the minimum FPR are shown below (Table 3). The test gives $\chi^2 = 886.5033$ which equates to a probability very close to 0: the classifier deviates from chance at the $p < 0.01$ significance level which means the decision classes are dependent on repeated episodes.

Table 3. Classifier results at maximum accuracy. Rno predicts no repeat episodes and Uno predicts repeat episodes.

<table>
<thead>
<tr>
<th></th>
<th>Uno</th>
<th>Rno</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeat</td>
<td>407</td>
<td>170</td>
<td>577</td>
</tr>
<tr>
<td>No Repeat</td>
<td>7</td>
<td>945</td>
<td>952</td>
</tr>
<tr>
<td>Total</td>
<td>414</td>
<td>1115</td>
<td>1529</td>
</tr>
</tbody>
</table>

The most encouraging feature is that the unreliable intention decision has very few errors: only seven out of 407, which means it is not generating false alarms that would dilute the effect of warning practitioners. Although the assessments classified as reliable have many more errors, where people go on to repeat, the status quo is to consider all assessments as reliable. Ideally, we would like to move as many as possible of these into the unreliable class but, for now, the real impact of the research is in demonstrating that unreliability can be meaningfully predicted so that practitioners take the warning seriously.

Conclusion
The results show that assessments where the answer to current intention of suicide is “no” can be divided into two classes, unreliable and reliable negative intention. A non-standard method was used due to the amount of missing data in the samples but future work would explore more conventional methods such as regression, support vector machines, and decision trees, for example. Either way, the classification is based on other cues collected during an assessment, which can be used to develop a reliability measure for current intention that smooths the influence of intention on the clinical risk judgement. Instead of having a categorical “no intention”, a graded value for intention
could be generated where some negative intention statements could even be regarded as the opposite. Alternatively, the classification based on reliability may be incorporated as a risk factor on its own, or incorporated in some of the cues that are already included in the risk assessment.

Alerting assessors to unreliability of intention would, in itself, help improve the data collection process, by ensuring assessors pay attention to issues that would not normally be considered for low-risk people. The top seven cues (Table 1) used for our reliability measure have a significant percentage of missing data and our results strongly support the need for them to be collected as a matter of course. However, they are all related to previous history of suicide and the next step is to determine whether a suitable reliability measure can be found for people without a history. The non-history cues from Cue 9 (Table 1) suggest this will be possible because they are able to separate the classes; a reliability indicator is needed because the table shows the cues are missing from a high percentage of assessments and should certainly be collected for those patients without a history of suicide. The upshot will be a measure applicable for all populations that will increase assessors’ understanding of suicide risk and help prevent future attempts.

Acknowledgement

This work was part supported by Grant SRG-0-060-11 awarded to C.D. Buckingham from the American Foundation for Suicide Prevention. The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Foundation for Suicide Prevention.

References

ABSTRACT

Semantic role labeling (SRL), which extracts shallow semantic relation representation from different surface textual forms of free text sentences, is important for understanding clinical narratives. Since semantic roles are formed by syntactic constituents in the sentence, an effective parser, as well as an effective syntactic feature set are essential to build a practical SRL system. Our study initiates a formal evaluation and comparison of SRL performance on a clinical text corpus MiPACQ, using three state-of-the-art parsers, the Stanford parser, the Berkeley parser, and the Charniak parser. First, the original parsers trained on the open domain syntactic corpus Penn Treebank were employed. Next, those parsers were retrained on the clinical Treebank of MiPACQ for further comparison. Additionally, state-of-the-art syntactic features from open domain SRL were also examined for clinical text. Experimental results showed that retraining the parsers on clinical Treebank improved the performance significantly, with an optimal F1 measure of 71.41% achieved by the Berkeley parser.

INTRODUCTION

Natural language processing (NLP) technologies are important for unlocking information embedded in narrative reports in electronic health record (EHR) systems. Although various NLP systems have been developed to support a wide range of computerized medical applications, such as bio-surveillance and clinical decision support, extracting semantically meaningful information from clinical text is still a challenge.

In the biomedical domain, semantic relation extraction systems, such as LSP, MedLEE, MedEx for clinical text and SemRep for biomedical literature, have shown good performance and been widely used in different applications. These early-stage systems were often based on manually extracted patterns, following the sub-language theory. According to the sub-language theory, the language of a closed domain (e.g., medicine and biomedicine) has special syntactic patterns as well as a limited number of main semantic types. Therefore, possible semantic relations could be identified by restricted constraints of syntactic and/or semantic patterns. However, a careful examination of syntactic alterations that express the same semantic relations in biomedical text reveals that even in a semantically restricted domain, syntactic variations are common and diverse. Thus, the coverage and scalability of manually extracted patterns may not be sufficient for those syntactic variations. In recent years, promoted by increasing challenges held by different portals (e.g., BioCreative, BioNLP, i2b2 and SemEval), more and more automatic information extraction systems have been built for different biomedical subdomains using data-driven statistical methods, such as machine learning algorithms. However, diverse syntactic variations still remain as an essential problem to extract semantic information from biomedical text, especially for clinical text, which contains more fragments and ill-formed grammars.

Figure 1. A syntactic parse tree with semantic roles added (ARGs).
One potential solution to this problem is semantic role labeling (SRL) (also known as shallow semantic parsing), which focuses on unifying variations in the surface syntactic forms of semantic relations. Specifically, the task of SRL is to label shallow semantic relations in a sentence as predicate argument structures (PAS). A predicate usually refers to a word indicating an event or a relation, and arguments (ARGs) refer to syntactic constituents representing different semantic roles in the event or relation. For each predicate, arguments representing the most important semantic roles are labeled with numbers, usually from ARG0 to ARG5. In addition, arguments representing modifiers of events (i.e., location, time, manner, etc.) are labeled as ARGMs. Taking the sentence "She should decrease the prednisone by 1-mg weekly" in Figure 1 as an example, the verb phrase "decrease" is the predicate indicating the event; the noun phrase "She" represents the role of ARG0, indicating the initiator/executor of the action "decrease"; the noun phrase "the prednisone" represents the role of ARG1, indicating the receptor of the action "decrease" (i.e. the entity decreased); while the prepositional phrase "by 1-mg weekly" represents the manner of how to decrease the prednisone (ARGM-Manner).

Shallow semantic relations, or PASs are usually applied as features for machine learning algorithms, sentence structural representations in kernel-based models or inference rules in different applications, including question answering, text summarization and information extraction, etc. Especially, PASs have been investigated in various biomedical sub-domains and made positive contributions in semantic information extractions, such as extracting drug-drug interactions from biomedical literature and temporal relations from clinical text.

Generally, a typical SRL system is built by using machine-learning methods based on annotated corpora. Since semantic roles are formed by syntactic constituents, two corpora are needed to build SRL systems, namely a corpus of syntactic parse trees and a corresponding corpus of semantic roles annotated on it. The most widely used large-scale corpora in open domain are the Penn Treebank and the SRL corpus PropBank developed on it. Many state-of-the-art syntactic parsers have been developed and applied to SRL in open domain. Some previous studies attempted to adapt these parsers (e.g., the Stanford Parser) to clinical text using medical lexicons. Recent years have also seen emerging efforts for syntactic annotation guidelines and corpora of clinical text. For example, the MiPACQ corpus (a multi-source integrated platform for answering clinical questions) annotated syntactic trees for 13,091 sentences following the Penn Treebank Style. Furthermore, several SRL corpora were developed for clinical text following the PropBank Style. The available corpora were of different genres and note styles, including operative notes, radiology notes from the SHARP Area 4 project (Strategic Health IT Advanced Research Projects), colon cancer pathology and clinical notes from the MiPACQ corpus and the THYME corpus (Temporal Histories of Your Medical Events). Based on those corpora, studies have been conducted to investigate SRL techniques for clinical text from EHRs. Albright et al. (2013) and Zhang et al. (2014) developed SRL systems on the MiPACQ corpus using dependency parse trees and constituent parse trees, respectively. Wang et al. (2014) built a SRL system on operative notes using an adapted parser.

Given that semantic roles are formed by syntactic constituents in the sentence, an effective parser to first recognize those syntactic constituents is critical for developing a practical SRL system. Furthermore, an effective feature set to describe the syntactic patterns between the predicate and the argument is also essential to SRL. Although previous works have compared different syntactic parsers and representations for biomedical event extraction from literature, there are no formal evaluations and comparisons of state-of-the-art parsers and features for SRL in the medical domain.

In this study, we evaluated the SRL performance of three state-of-the-art constituent syntactic parsers: the Stanford parser, the Charniak parser and the Berkeley parser, using the MiPACQ corpus. We focused on constituent parse trees here because they could be directly converted to dependency parse trees. The purposes of this study were two-fold: (1) to evaluate the SRL performance of existing state-of-the-art English parsers on clinical text, both the original parsers developed on Penn Treebank and parsers retrained on the clinical Treebank were examined; and (2) to validate the effectiveness of state-of-the-art syntactic features for SRL in the open domain on clinical text. To the best of our knowledge, this is the first comprehensive study that investigates the influence of syntactic parsing and features for SRL on clinical text using multiple state-of-the-art parsers.
METHODS

Dataset
This study used the MiPACQ dataset for SRL experiment. MiPACQ is built from randomly selected clinical notes and pathology notes of Mayo Clinic related to colon cancer. Layered linguistic information is annotated in MiPACQ, including part of speech (POS) tags, syntactic Treebank, PASs for SRL, named entities, and semantic information from Unified Medical Language System. The syntactic Treebank annotations in MiPACQ follow the Penn Treebank guidelines, and the predicate-argument structure annotations for SRL follow PropBank guidelines. 13,091 sentences are annotated with syntactic trees. Among them, 6,145 sentences in MiPACQ are annotated for SRL, including 722 verb predicates with 9,780 PASs and 415 nominal predicates with 2,795 PASs.

The basic SRL system

Figure 2. Study design for semantic role labeling of clinical text

Figure 2 shows the study design for SRL of clinical text. Basically, the SRL system can be partitioned into the training stage and the testing stage. In the training stage, gold-standard syntactic trees of the training data set annotated in MiPACQ are used for feature extraction. A SRL task consists of two sub-tasks, the argument identification sub-task and the argument classification sub-task. First, a binary non-Argument vs. Argument classifier is built as the argument identifier on the entire dataset for all the predicates, instead of building one model per predicate. For argument classification, a multi-class classifier is built to assign semantic roles to arguments of all the predicates. In the testing stage, syntactic trees automatically generated by the syntactic parser are used for feature extraction. For each predicate, the argument candidates first go through the argument identifier. If one candidate is identified as an argument, it will go through the argument classifier that assigns the semantic role.

Comparing syntactic parsers and features

Three widely used state-of-the-art syntactic parsers, the Stanford parser, the Charniak parser, and the Berkeley parser are investigated for their influence on the SRL performance in our study. Moreover, state-of-the-art features, most of which are syntactic features, commonly used in open domain and biomedical domain are extracted and compared for use with clinical text.

Syntactic Parsers
Stanford parser: Stanford parser was initially developed based on un-lexicalized probabilistic context-free grammar (PCFG). The lexicalized PCFG parser and the dependency parser shared a factored product model, where the preference of PCFG phrase structures and lexical dependencies are combined by efficient exact inference, using an A* algorithm.
Charniak parser\textsuperscript{24}: Charniak parser is a high-performance lexicalized parser. It achieved an $F_1$ measure of 91.0\% on the Penn Treebank. This parser is constructed by re-ranking the 50-best parse trees generated from a coarse-to-fine generative parser, using a discriminative Maximum Entropy model.

Berkeley parser\textsuperscript{25}: Berkeley parser is also based on PCFG. In addition to local correlations between syntactic tree nodes, this parser learns to capture the non-local correlations between tree nodes using an EM algorithm based model.

Features

Similar to previous work of SRL for biomedical literature\textsuperscript{41,19} and clinical text\textsuperscript{35}, we adopted the common features used in current state-of-the-art SRL systems. The features include baseline features from the original work of Gildea and Jurafsky (2002)\textsuperscript{42}, advanced features taken from Pradhan et al. (2005)\textsuperscript{10} and feature combinations from Xue and Palmer (2004)\textsuperscript{26}.

The features can be categorized into three major groups: (1) basic features include the lexical and syntactic features of the predicate and the argument; (2) context features include features of the syntactic nodes surrounding and the syntactic paths between the predicate and the argument; (3) feature combinations are feature tuples formed of two unitary features from the previous two groups. The complete feature set is described in Table 1. Except for the lemma of the predicate word and the relative position between the argument and the predicate, all the rest of the features are at the syntactic level and need to be extracted from the parse tree. For a more clear illustration, Table 2 lists the specific features extracted for the argument candidate “1-mg weekly” of the predicate “decrease” in the example sentence shown in Figure 1.

Table 1. Feature list of semantic role labeling

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>basic features</strong></td>
<td></td>
</tr>
<tr>
<td>Predicate</td>
<td>Lemmatization of the predicate word</td>
</tr>
<tr>
<td></td>
<td>Voice of the verb predicate, i.e., active or passive</td>
</tr>
<tr>
<td>Argument</td>
<td>Syntactic head, first word, last word of the argument phrase and their POS tags</td>
</tr>
<tr>
<td></td>
<td>Syntactic category of the argument node</td>
</tr>
<tr>
<td></td>
<td>Whether the argument is a preposition phrase</td>
</tr>
<tr>
<td></td>
<td>Enriched POS of prepositional argument nodes (e.g., PP-for, PP-in)</td>
</tr>
<tr>
<td>Relative position</td>
<td>Relative position of the argument with respect to the predicate (before or after)</td>
</tr>
<tr>
<td><strong>Context features</strong></td>
<td></td>
</tr>
<tr>
<td>Production rule of predicate</td>
<td>Production rule expanding the predicate parent node</td>
</tr>
<tr>
<td>Syntactic category of argument neighbors</td>
<td>Syntactic categories of the parent, left sister and right sister of the argument node</td>
</tr>
<tr>
<td>Path</td>
<td>Syntactic path linking the predicate and an argument</td>
</tr>
<tr>
<td>No-direction path</td>
<td>Like Path, but without traversal directions</td>
</tr>
<tr>
<td>Partial path</td>
<td>Path from the argument to the lowest common ancestor of the predicate and the argument</td>
</tr>
<tr>
<td>Syntactic frame</td>
<td>Position of the NPs surrounding the predicate</td>
</tr>
<tr>
<td><strong>Feature combinations</strong></td>
<td>Predicate and head word of the argument</td>
</tr>
<tr>
<td></td>
<td>Predicate and Syntactic category of the argument</td>
</tr>
<tr>
<td></td>
<td>Predicate and relative position</td>
</tr>
<tr>
<td></td>
<td>Predicate and path</td>
</tr>
</tbody>
</table>

Experiments

PASs with at least one argument were used for the experiment. We used the open source toolkit, Liblinear\textsuperscript{43} as implementations of the support vector machine algorithm. For each implemented method, all parameters were tuned for optimal performance.

Experiments and systematic analysis were conducted as follows:
1. Evaluate SRL performance of parsers with their default settings: In this experiment, we directly applied the three parsers to process all sentences of the test dataset. All the parsers were invoked with their default settings and models, which had been trained on the Penn Treebank.

### Table 2. An example of features extracted for semantic role labeling.

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Feature value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic features</strong></td>
<td></td>
</tr>
<tr>
<td>Predicate</td>
<td>decrease</td>
</tr>
<tr>
<td>Argument</td>
<td>hw_1-mg, hw_pos_NN, fw_by, fw_pos_IN, lw_weekly, lw_pos_RB PP Yes PP-by</td>
</tr>
<tr>
<td>Relative position</td>
<td>after</td>
</tr>
<tr>
<td><strong>Context features</strong></td>
<td></td>
</tr>
<tr>
<td>Subcategory of predicate</td>
<td>VP→VB–NP–PP</td>
</tr>
<tr>
<td>Syntactic category of argument neighbors</td>
<td>scp_VP, scl_NP, scr_null</td>
</tr>
<tr>
<td>Path</td>
<td>PP↑VP↓VB</td>
</tr>
<tr>
<td>No-direction path</td>
<td>PP_VP_VB</td>
</tr>
<tr>
<td>Partial path</td>
<td>PP↑VP</td>
</tr>
<tr>
<td>Syntactic frame</td>
<td>Position of the NPs surrounding the predicate</td>
</tr>
<tr>
<td>Feature combinations</td>
<td>decrease_1-mg, decrease_PP, decrease_after, decrease_VB↑VP↓PP</td>
</tr>
</tbody>
</table>

* In the syntactic path feature, ↑ indicates a link from a child node to its parent, and ↓ indicates a link from a parent node to its child.

2. Evaluate SRL performance of parsers re-trained on the clinical Treebank: To assess if the annotation of clinical Treebank could improve the performance of SRL, we applied three parsers retrained on the MiPACQ Treebank. We conducted ten-fold cross validation evaluation for each parser. The cross-validation involved dividing the clinical corpus equally into 10 parts, and training the parser on 9 parts with testing on the remaining part each time. We repeated the same procedure 10 times, one for each part, and then combined the results from the 10 parts to report the performance.

3. Evaluate SRL performance of each syntactic feature: To validate if syntactic features commonly used in the open domain were effective for clinical text, we conducted multiple runs of experiments, adding one new syntactic feature into the feature set for each run. The experimental results were compared to check the effectiveness of each feature.
Evaluation

Precision (P), recall (R) and $F_1$-measure ($F_1$) were used as evaluation metrics for argument identification (AI) and combined SRL task. Precision measures the percentage of correct predictions of positive labels made by a classifier. Recall measures the percentage of positive labels in the gold standard that were correctly predicted by the classifier. $F_1$-measure is the harmonic mean of precision and recall. During the process of argument classification (AC), the boundaries of candidate arguments are already identified by the argument identification step. Therefore, the accuracy (Acc) of the classifier was used for evaluation, which is defined as the percentage of correct predictions with reference to the total number of candidate arguments correctly recognized in the argument identification step. Ten-fold cross validation was employed for performance evaluation.

RESULTS

Table 3 illustrates the performance of semantic role labeling systems, which were trained on the gold standard syntactic trees and tested on the parsing results of Stanford, Charniak and Berkeley, as well as the gold standard syntactic trees, respectively. For these experiments, the whole feature set described in Table 2 was used. The original parsers trained on the Penn Treebank produced relatively lower performance. Charniak got the lowest $F_1$-measure of 61.40%, whereas Berkeley outperformed the other two parsers with a $F_1$-measure of 68.15%. After retraining on the clinical Treebank, the performance of all three parsers increased significantly, with the optimal $F_1$-measure of 71.38% achieved by Berkeley. Testing on the gold standard parse trees yielded a $F_1$-measure of 82.13%.

Table 3. Performance of semantic role labeling systems trained on the gold standard syntactic trees and tested on the parsing results of Stanford, Charniak and Berkeley, and the gold standard syntactic trees, respectively (%)

<table>
<thead>
<tr>
<th>Parser</th>
<th>Model</th>
<th>AI</th>
<th>AC</th>
<th>AI+AC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>$F_1$</td>
</tr>
<tr>
<td>Stanford</td>
<td>Original</td>
<td>70.42</td>
<td>82.17</td>
<td>75.84</td>
</tr>
<tr>
<td></td>
<td>Retrained</td>
<td>75.74</td>
<td>85.27</td>
<td>80.22</td>
</tr>
<tr>
<td>Charniak</td>
<td>Original</td>
<td>67.75</td>
<td>74.54</td>
<td>70.98</td>
</tr>
<tr>
<td></td>
<td>Retrained</td>
<td>74.20</td>
<td>87.09</td>
<td>80.12</td>
</tr>
<tr>
<td>Berkeley</td>
<td>Original</td>
<td>72.88</td>
<td>83.09</td>
<td>77.64</td>
</tr>
<tr>
<td></td>
<td>Retrained</td>
<td>76.72</td>
<td>85.37</td>
<td>80.81</td>
</tr>
<tr>
<td>Gold Standard</td>
<td></td>
<td>91.41</td>
<td>91.60</td>
<td>91.51</td>
</tr>
</tbody>
</table>

* AI: argument identification  AC: argument classification

To investigate whether syntactic features commonly used in the open domain are also effective for clinical text, multiple experiments were conducted by adding one new syntactic feature incrementally for each run. Table 4 lists the SRL performance of both the gold standard corpus and the parse results of the retrained Berkeley. As the baseline, the first run adopted all the basic features of predicate, argument and their relative position. Numbers in parenthesis show the changes to $F_1$-measure of argument identification and accuracy of argument classification by adding each new feature. As illustrated in Table 4, all the syntactic features effective in the open domain were also helpful for argument identification of clinical text. The $F_1$-measure was improved consistently from 20.07% and 17.47% to 91.51% and 80.81% for the gold standard corpus and the retrained Berkeley parser, respectively. In addition to the basic features, phrase types of argument neighbors, and the three path features made the most contribution to argument identification. In contrast, for argument classification, the basic features already yielded an accuracy of 86.74% for the gold standard corpus and an accuracy of 83.78% for the retrained Berkeley. Since the path features between the predicate and an argument dropped the accuracy slightly, we conducted additional experiments by removing those features for argument classification, which improved the overall $F_1$-measure of our SRL systems to 82.14% (vs. 82.13%) for the gold standard corpus and to 71.41% (vs. 71.38%) for the retrained Berkeley parser.
**DISCUSSION**

Effective syntactic parsers and features are critical to establish a practical SRL system. This study takes the initiative to make a formal evaluation and comparison of SRL performance on a clinical text corpus MiPACQ, using three state-of-the-art syntactic parsers and common syntactic features used in open domain. Experimental results demonstrate that retraining parsers on clinical corpora could improve the SRL performance significantly, with an optimal F1-measure of 71.41% achieved by the Berkeley parser. Despite the telegraphic type of clinical text, state-of-the-art syntactic features in open domain also proved to be effective for clinical text.
In terms of SRL errors caused by syntactic parsers, a major category was that the parsers did not recognize a large number of syntactic constituents acting as arguments (Original Stanford: 1175, Charniak: 1377, Berkeley: 1262). Nevertheless, retraining parsers on the clinical Treebank reduced such errors greatly (Retrained Stanford: 887, Charniak: 973, Berkeley: 816). Another major type of syntactic problems that caused SRL errors was the essential syntactic structure ambiguities. For example, the sentence “He continues to note the sensation of bilateral leg numbness and pins and needle sensation with walking” contains conjunctive structures linking two phrases “the sensation of bilateral leg numbness”, and “pins and needle sensation”. It’s hard to determine if the prepositional phrase “with walking” only modifies the “pins and needle sensation” or both phrases.

Despite the unique characteristics of clinical text, such as fragments and ill-formed grammars, all the state-of-the-art syntactic features in the open domain contributed positively to clinical text, except for path features that dropped the accuracy of argument classification slightly. One possible reason for the decreased performance is that the specific semantic role of an argument in clinical text is dependent not only on syntactic paths but also on the clinical lexicon and relations. As an example, in the phrase “an advanced breast cancer treated with radiation therapy”, “an advanced breast cancer” is annotated as ARG2 (illness or injury) in the gold standard. However, it was mistakenly labeled as ARG1, because the extracted syntactic path features were similar to those of ARG1 in the corpus.

Our study has the following limitations. In this study, only the SRL performance of pathology notes and clinical notes are investigated. In the future, we plan to extend this study to other types of clinical notes such as operative notes, to assess the generalizability of our findings. In addition to parsers retrained on clinical text, parsers adapted to clinical text using domain lexicons and grammars could also be explored for SRL. Although this study mainly focused on syntactic features, existing clinical knowledge such as semantic types and relations will be further examined as features for SRL in the next stage. Besides, there are other widely used parsers in open domain such as OpenNLP, LingPipe and Gate in addition to the three parsers explored in this study. An extension of our study to other parsers will be conducted in the further, so that to reach the optimal SRL performance on clinical text.

**CONCLUSION**

This study made a formal evaluation and comparison of SRL performance on a clinical text corpus, MiPACQ, using three state-of-the-art parsers, the Stanford parser, the Berkeley parser, and the Charniak parser and state-of-the-art syntactic features from the open domain. Experimental results validated the effectiveness of retraining parsers with a clinical Treebank, with an optimal F1-measure of 71.41% achieved by the Berkeley parser. The results also demonstrated that common syntactic features in open domain could contribute positively to the clinical text.

**ACKNOWLEDGEMENT**

This study was supported by grants from the NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282. We would like to thank the MiPACQ team for the development of the corpora used in this study.

**References**

22 Kingsbury P, Palmer M. From TreeBank to PropBank. LREC’02; 2002.
35 Zhang YT, Buzhou; Jiang, Min; Wang, Jingqi; Wu, Yonghui; Xu, Hua. domain adaptation for semantic role labeling of clinical text. AMIA; 2014; Washington DC, USA; 2014.
36 Wang YP, Serguei; Ryan, James O.; Melton, Genevieve B. Semantic Role Labeling for Modeling Surgical Procedures in Operative Notes. AMIA; 2014; Washington DC., USA; 2014.
44 Jiang MH, Yang; Fan, Jung-wei; Tang, Buzhou; Denny, Josh; Xu, Hua. Parsing clinical text: how good are the state-of-the-art parsers? BMC Medical Informatics and Decision Making. 2015.
A Decade of Experience in Creating and Maintaining Data Elements for Structured Clinical Documentation in EHRs

Li Zhou, MD, PhD\textsuperscript{1,2,3}, Sarah Collins, RN, PhD\textsuperscript{1,2,3}, Stephen J. Morgan, MD\textsuperscript{1,2,3}, Neelam Zafar, MD, MHA\textsuperscript{1}, Emily J. Gesner DNP, RN-BC, Martin Fehrenbach\textsuperscript{4}, Roberto A. Rocha, MD, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Clinical Informatics, Partners HealthCare System, Boston, MA; \textsuperscript{2}Brigham and Women’s Hospital, Boston, MA; \textsuperscript{3}Harvard Medical School, Boston, MA; \textsuperscript{4}Institute of Medical Biometry and Informatics, Heidelberg University, Germany

Abstract

Structured clinical documentation is an important component of electronic health records (EHRs) and plays an important role in clinical care, administrative functions, and research activities. Clinical data elements serve as basic building blocks for composing the templates used for generating clinical documents (such as notes and forms). We present our experience in creating and maintaining data elements for three different EHRs (one home-grown and two commercial systems) across different clinical settings, using flowsheet data elements as examples in our case studies. We identified basic but important challenges (including naming convention, links to standard terminologies, and versioning and change management) and possible solutions to address them. We also discussed more complicated challenges regarding governance, documentation vs. structured data capture, pre-coordination vs. post-coordination, reference information models, as well as monitoring, communication and training.

Introduction and Background

Structured clinical documentation is an important component of electronic health records (EHRs) and plays an important role in clinical care (e.g. dependency for clinical decision support [CDS]), administrative functions (e.g. extracted to support billing, quality assessment and reporting), research activities, and other areas. \textsuperscript{1,2} When well implemented in the context of the clinical workflow, structured documents can save clinicians’ time, assure professional practice standards and clinical thoroughness, and may reduce potential medical errors through clinical decision support interventions. \textsuperscript{2}

Different levels of information aggregation for structured documents are shown in Figure 1. Data elements serve as basic building blocks for composing document templates that are used for generating clinical documents (e.g. provider notes and forms). Appropriate underlying terminologies and information models used to assemble these data elements, closely correlate to the quality of the produced documents, and are keys to a successful implementation. The management of these data elements remains a critical and demanding issue for EHR systems.

Researchers have made significant efforts in developing methodologies for specifying the structure and semantics of clinical documents. Standard development organizations, professional associations and other healthcare organizations have developed standards and models in this area. A few known efforts include: 1) HL7 Clinical Document Architecture (CDA), \textsuperscript{3} which is a document markup standard and specifies an exchange model for clinical documents; 2) HL7 Reusable Information Constraint Templates, \textsuperscript{4} which are used to constrain the structures of a portion of atomic concepts, such as a laboratory report in a CDA document; 3) OpenEHR Archetype Model, \textsuperscript{5} which consists of a domain-level definition in the form of archetypes and templates and allows clinical experts to be able to structure their own data in the way they require it; 4) Intermountain Health Care’s Detailed Clinical Models\textsuperscript{6-9} whose core building block is a “clinical element” represented in a recursive data structure for capturing detailed clinical information; and 5) recent Clinical Information Modeling Imitative by HL7 that aims to create a shared repository of detailed clinical information models and binds the models to standard terminologies. \textsuperscript{4} EHR vendors most commonly use proprietary approaches to structured documentation and template management.

In this paper, we present our experience in creating and managing data elements for different EHR systems, focusing mainly on lessons learned from two legacy systems (a homegrown ambulatory EHR system and a commercial inpatient documentation system) and a newly implemented commercial EHR system. In particular, we
use a basic but nontrivial domain, flowsheet data elements, as an example to illustrate and discuss issues and challenges we encountered.

Case Study Site

Partners Healthcare System is an integrated health care system in the Boston area, founded in 1994 by Brigham and Women’s Hospital and Massachusetts General Hospital. It also includes several community and specialty hospitals, community health centers, and other health-related entities. The Longitudinal Medical Record (LMR) is an internally developed, full-featured, and Meaningful Use-Certified EHR, including primary care and subspecialty semi-structured notes, orders, problem lists, medication lists, allergies, laboratory tests, clinical decision support, quality reporting and other functionalities. It was developed and implemented in early 2000s and used across Partners healthcare network. In 2007, Partners carried out the Acute Care Documentation (ACD) project that aimed to develop highly structured clinical content for the inpatient setting using a vendor system. In 2012, Partners cancelled the ACD project and announced a new initiative, known as Partners eCare, to implement a commercially available integrated EHR system at all its sites. On May 30, 2015, the first site, including Brigham and Women’s Hospital, Partners Home Care, and Dana Farber Cancer Institute, went live with the new EHR system as part of the Partners eCare initiative. In the following, we describe three case studies based on our experience in creating and maintaining data elements for these EHR systems.

Case Study 1: the Longitudinal Medical Record (LMR) System

A Brief History

LMR allows great flexibility for users to create their own data elements and templates at a local level based on their needs. At the early stages of the systems use, data elements were created on an ad-hoc manner based on the specific needs of individual practices and providers and they were generally not shared among users. Because the granularity (level of detail) and presentation of the data elements change over time and vary by domain, by practice, or by user, their reusability is low. The initial design of these data elements applied the “concept-attribute-value (or section-question-answer)” paradigm. Figure 2 demonstrates an example of a data element for lung symptoms developed by a local oncology team. This type of design mixed data and presentation, compromising reuse and data consistency.

In an effort to address these problems, in 2006, the LMR development team and the Partners Knowledge Management (KM) team launched a project that aimed to maximize re-use of data elements and templates, and to support CDS, pay for performance, and quality reporting. A few solutions were proposed and implemented: 1) build an editor that facilitates controlled authoring of data elements and templates; 2) define a shared data element dictionary to manage all elements needed for composing structured clinical documents, where users can create their own data elements or search existing ones; and 3) creation of a limited set of “enterprise” data elements that were flagged and locked (i.e. editable only by authorized KM team members). These solutions did not entirely solve the fundamental limitations of the system. For example, the search function of the editor was relatively simple. LMR Analysts preview existing data elements and templates for new clinical practices and customize as needed. If no existing elements and templates for a specific practice type were found, new ones were promptly created from scratch. In addition, users often expected a high level of customization. Overtime, many similar, or sometimes identical, concepts were added into the data element dictionary. For example, more than 127 same or similar “alcohol use” data elements were created. Compared to a total of 28,400 data elements in 2007, the total number has increased 2.2 times within 3 years to 62,600 data elements in 2010, and continued to grow dramatically in the following years. We coined this phenomenon as the uncontrolled replication of “cancerous” data elements.

Another significant challenge we faced was that these data elements represented concepts from diverse clinical domains, often with duplicate data elements for the same clinical concepts existing and in use in different EHR modules (e.g., problems, medications, procedures, laboratory tests, assessments, and other clinical concepts). For example, a separate “hemoglobin A1c” data element was created specifically for a health maintenance template, but it was not necessarily or directly linked to a patient’s laboratory test or result. It was difficult to define an appropriate scope for a separate set of data elements that are only used in clinical notes and forms, what terminologies (if available) should be used to encode them, and how to link these data elements to other concepts within an EHR.
An Example: Flowsheet Data Elements

Here we use flowsheet data elements as an example to demonstrate in more detail the important issues related to the definition, implementation, and management of clinical data elements. EHR flowsheets allow clinicians to track specific elements of a patient’s health over time and are helpful in identifying trends within these data. As a common method of documenting clinical observations and physiological measurements, flowsheets are an important component of EHRs and are used as a data source for clinical decision support, safety and quality reporting, clinical trials eligibility criteria, and overall data sharing. When well aligned with the user’s clinical workflow, flowsheets can save time and promote consistent documentation, potentially reducing medical errors. However, studies have highlighted the challenges with designing an efficient flowsheet user interface, maintaining the flowsheet ontology, adopting a proper terminology, and measuring compliance in clinician documentation.

Some of the most common data elements of an EHR flowsheet are vital signs, such as height, weight, and blood pressure. In different clinical settings and specialties, flowsheets can be configured to reuse standardized content (such as vital signs and pain scales consistent with quality measures and reportable outcomes), as well as to record more complex data elements. Flowsheets are inherently flexible and analogous to a spreadsheet organized with data element rows and columns that indicate the date and time of recorded observations. This flexibility allows the design and development of flowsheet data elements and templates to meet specific needs of multiple providers, but requires consideration for how data can be standardized and shared among all users of the EHR system. The composite nature of flowsheet data elements and templates requires that the semantic relationships between each component be clearly defined. Figure 3 (left) shows an example of the editor for a flowsheet data element “temperature.” In LMR, each data element can be defined with an identifier, name, abbreviation, result type, and at most two attributes. Each attribute can have a list of value options. Figure 3 (right) shows an example of flowsheet data entry form using a template composed by combining a set of flowsheet data elements. Figure 4 shows a patient’s flowsheet records in a chronological view.

We performed the following analyses on LMR flowsheet data elements. First, we looked at the total number of flowsheet related data elements and found 3,923 in total. Of these, only 36% had distinct names and the remaining were duplicates (exact same name) of existing data elements. The KM team and the Clinical Quality Reporting team at Partners have identified a set of data elements (by grouping data elements with the same meaning) for the purpose of quality reporting, pay for performance reporting, and CDS. We identified that these important data elements accounted for only less than 3% of the total.

Duplicate or cloned data elements have been frequently created in the LMR dictionary. For example, we found 57 data elements for blood pressure, 55 for weight, 42 for pulse, 21 for height, 15 for temperature, 12 for oxygen
saturation, 6 for respiratory rate, and 5 for pain level (0-10). Using weight as an example, 13 of the 55 data elements are synonyms and 2 are abbreviations. The remaining weight data elements were defined using different combinations of attributes (e.g., unit and measure method) and data type (e.g., string or numeric). Table 1 demonstrates the diversity of attributes or properties and their value sets for some of the flowsheet data elements.

In addition, even with customized structured data elements, we found that free-text comments were often used (sometimes as a workaround) in flowsheet data entries. Collins et al found that nurses use free-text comments in flowsheets as a method to communicate concerning events to physicians, and that these data were associated with survival outcomes of cardiac arrest patients. These phenomena needs to be further studied to provide better understanding of current limitations of EHR flowsheet data elements for automated analyses.

Table 1: Examples of Flowsheet Data Elements and Their Attributes and Value Set Found in LMR

<table>
<thead>
<tr>
<th>Data Elements</th>
<th>Attributes or Properties and Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure</td>
<td>Body location: right arm, left arm, right leg, left leg, both arms, wrist  Method: doppler, manual, auto  Position: sitting, standing, lying  Device: large cuff, small cuff, thigh cuff, pedi cuff, adult cuff, dinamap  As per: home monitoring, from note, doctor repeated, patient reported, nurse verified  Patient state: after relaxation, after exam, after medication, orthostatic  Comments: e.g. patient did not take medication, second reading</td>
</tr>
<tr>
<td>Pulse</td>
<td>Body location: radial, apical, brachial, carotid, peripheral, pedal  Method: oximeter, EKG, manual  Position: standing, sitting, lying  Rhythm: regular, irregular, abnormal  Abnormal Rhythm: murmur, pre-mature beats, sinus arrhythmia  As per: reported by nurse, observed, verified by doctor  Negation: no</td>
</tr>
<tr>
<td>Temperature</td>
<td>Body location: axillary, rectal, per ear  Patient condition: eating or drinking (hot or cold), post medication  As per: reported by nurse  Comments: e.g. patient eating chewing gum</td>
</tr>
<tr>
<td>Respiratory rate</td>
<td>Rhythm: regular, irregular, deep, rapid, shallow  Patient condition: pre and post nebulizer, coughing, respiratory stress, wheezing  Patient state: at rest  As per: verified by nurse</td>
</tr>
</tbody>
</table>

Case Summary

As showed above, the underlying design of data elements was rather simple in LMR (i.e., concept-attribute-value format). One of the original objectives was that by using a relatively simple structure it would allow a high level of flexibility for users to create and customize data elements and templates, which would lead to a wide adoption of structured documentation. However, at the early stage, well-defined data governance strategies were not established. There was no formal model or mechanism for structuring data elements, nor an efficient editing tool. Similarly, linkage and mapping to standard terminologies were not created for these elements. As a result, there is redundancy, inconsistency, minimal reusability, and a lack of interoperability of the existing forms and templates, resulting in continuously increasing maintenance costs.

Case Study 2: the Acute Care Documentation (ACD) Project

A Brief History

Partners Healthcare Systems began the “Acute Care Documentation” (ACD) project in 2007 with the goal of configuring structured clinical documentation using a commercial EHR for implementation at BWH and MGH. The documentation templates were intended for use by nurses, physicians, physical therapists, dieticians and other health professionals that care for patients in the acute and critical care settings. The ACD project was ended just before implementation of the system, but the content creation had produced over 12,000 data elements that were used to create more than 1,200 documentation templates. Based on lessons learned from LMR, two major changes were implemented in the ACD project to control the 'neoplastic' behavior of data elements replication to represent clinical topics: 1) a structured naming convention and best practices for data element names, which help standardize structure and usage; 2) a software tool that allows users to create and customize data elements and templates, which helps to capture and share knowledge among clinicians.

Table 2. Counts of Vital Sign Data Elements

<table>
<thead>
<tr>
<th>Content Component (additional Text Search Term)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>127</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td>49</td>
</tr>
<tr>
<td>Systolic</td>
<td>14</td>
</tr>
<tr>
<td>Diastolic</td>
<td>12</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>37</td>
</tr>
<tr>
<td>Respiratory Rate</td>
<td>20</td>
</tr>
<tr>
<td>Alarm</td>
<td>24</td>
</tr>
<tr>
<td>hemodynamic</td>
<td>6</td>
</tr>
<tr>
<td>Heart Rhythm</td>
<td>0</td>
</tr>
<tr>
<td>Oxygen Saturation (O2)</td>
<td>52</td>
</tr>
<tr>
<td>Weight</td>
<td>31</td>
</tr>
<tr>
<td>Height</td>
<td>2</td>
</tr>
</tbody>
</table>

1296
practices for data definitions and reuse were defined upfront, and 2) a dedicated team of analysts were the only individuals creating data elements. These analysts were trained in using the defined structured naming conventions leveraging the International Standards Organization/International Electrotechnical Commission (ISO/IEC) 11179 standard. End-users and other clinical stakeholders helped define the documentation templates, but did not have any control in defining or customizing data elements.

An Example – Vital Sign Data Elements

We classified the ACD content in different content categories. A total of 4,188 data elements were classified into 13 categories. Counts for the top 3 categories were: 1,081 wound documentation data elements, 1,440 tube and drain related data elements, and 374 vital sign data elements. Table 2 shows the counts of data elements in the vital signs category. Despite the effort to manage data elements, duplicates and overlap were again present. Some duplicates were unintended and others were “by design” to accommodate system constraints in which a defined data element cannot handle multiple instances of that concept (e.g., multiple arterial blood pressures, multiple wounds). For example, “Blood pressure, diastolic arterial [mmHg]” and “Blood pressure, diastolic arterial 2 [mmHg]” are data fields for measured diastolic arterial blood pressure at a generic anatomical location. These represented essentially the same concept, yet the system requires 2 entries for tracking measurements overtime that take place in two different body locations. This phenomenon repeated for arterial and systolic blood pressure and likely was a factor in the perhaps unintended specification of “Blood pressure, diastolic femoral arterial [mmHg]” and “Blood pressure, diastolic femoral arterial 2 [mmHg]” given that “Blood pressure, diastolic arterial [mmHg]” was already defined. Table 3 shows a set of blood pressure data element names. With this approach a femoral measurement saved in the generic data field, would lead to inconsistent data retrieval.

Case Summary

Based on lessons learned from LMR document template creation, efforts were made to define data elements in the ACD project using a standard naming convention, but also to reuse data elements throughout the system. However, as described above, technical constraints of the system limited the ability to define content according to these identified best practices. This resulted in a data element dictionary with numerous “known” flaws. Such experiences re-confirmed the need to define practical approaches to improve the consistency and reuse of data elements within EHR systems.

Case Study 3: The Partners eCare Project (PeC)

A Brief History

Partners began the “Partners eCare Project” (PeC) in 2012 with the goal of a single patient record across the entire healthcare system using a different proprietary vendor EHR than the ACD project. The PeC project was significantly larger than LMR and ACD projects over a shorter period of time, greatly increasing the complexity of managing data elements. The PeC project went live with its first site on May 31, 2015 and has been in a process of successive implementations since that time. Due to certain functionality, the proprietary system separates structured data elements from flowsheet data elements. At the first went-live site, the total number of structured data elements that have been confirmed to be in use by end users is 15,209. The total number of flowsheet row data elements is 46,575. The project governance was structured based on the EHR system ‘modules’ that corresponded to a variety

<table>
<thead>
<tr>
<th>Table 3. Examples of ACD Blood Pressure Data Element Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial pressure left [mmHg]</td>
</tr>
<tr>
<td>Atrial pressure right mean [mmHg]</td>
</tr>
<tr>
<td>Blood pressure, diastolic arterial [mmHg]</td>
</tr>
<tr>
<td>Blood pressure, diastolic arterial 2 [mmHg]</td>
</tr>
<tr>
<td>Blood pressure diastolic femoral arterial [mmHg]</td>
</tr>
<tr>
<td>Blood pressure diastolic femoral arterial 2 [mmHg]</td>
</tr>
<tr>
<td>Blood pressure diastolic noninvasive [mmHg]</td>
</tr>
<tr>
<td>Blood pressure diastolic orthostatic lying [mmHg]</td>
</tr>
<tr>
<td>Blood pressure diastolic orthostatic sitting [mmHg]</td>
</tr>
<tr>
<td>Blood pressure diastolic orthostatic standing [mmHg]</td>
</tr>
<tr>
<td>Blood pressure four minute recovery [free text]</td>
</tr>
<tr>
<td>Blood pressure highest during intubation [free text]</td>
</tr>
<tr>
<td>Blood pressure lowest during intubation [free text]</td>
</tr>
<tr>
<td>Blood pressure mean arterial [mmHg] [free text]</td>
</tr>
<tr>
<td>Blood pressure mean arterial 2 [mmHg] [free text]</td>
</tr>
<tr>
<td>Blood pressure pre-procedure [free text]</td>
</tr>
<tr>
<td>Blood pressure Resting [free text]</td>
</tr>
<tr>
<td>Blood pressure seated rest [free text]</td>
</tr>
<tr>
<td>Blood pressure six minute recovery [free text]</td>
</tr>
<tr>
<td>Blood pressure supine rest [free text]</td>
</tr>
<tr>
<td>Blood pressure systolic arterial [mmHg]</td>
</tr>
<tr>
<td>Blood pressure systolic arterial 2 [mmHg]</td>
</tr>
<tr>
<td>Blood pressure systolic noninvasive [mmHg]</td>
</tr>
<tr>
<td>Blood pressure systolic orthostatic sitting [mmHg]</td>
</tr>
<tr>
<td>Blood pressure systolic orthostatic standing [mmHg]</td>
</tr>
<tr>
<td>Blood pressure two minute recovery [free text]</td>
</tr>
</tbody>
</table>
of clinical settings or specialty areas (e.g., inpatient acute care, emergency department, anesthesia, outpatient, and home health). Within each module, reuse and consistent naming of data elements was encouraged; however, reuse across modules was not emphasized as a priority. Problems noted were 1) duplication and overlap of data elements, 2) inconsistent naming convention, and 3) inconsistent data definitions for similar data concepts. During and after initial go-live, efforts were made to implement standardized and consistent data documentation as well as principles and best practices for customized or “local” data elements. We also developed a practical approach to creating reference models for clinical topics that have been determined to be high priority and that should be shared across applications within the EHR (e.g., pain and wound). This effort has been discussed in a previous publication

**An Example – Blood Pressure Documentation**

We extracted all existing data elements within the structured documentation forms and flowsheet records. We then used keyword search to identify data elements related to “blood pressure”. There are 4 data elements used on documentation forms that contained the word “blood pressure”, plus 40 others defined for flowsheets. Examples are shown in Table 4. The multiple data elements with overlapping definitions are partially due to the technical constraints of the system as well as the findings that the EHR configuration was based more upon end-user experience rather than consistent data definitions. The definition of data elements and configuration of the user interface was driven by large groups of subject matter experts (SMEs) and prioritized increasing usability and specialty-specific needs. More often than not, it is assumed that there was little to no search for similar data element definitions used in other modules prior to creating a new data element. The result is the overlapping data definitions that currently exist less than a year after going live with the first site.

**Case Summary**

The lessons learned from our previous experience with LMR and the ACD project did not sufficiently benefit our initial efforts configuring and implementing a vendor-based EHR system. Since those initial efforts, we have established a governance infrastructure and a set of principles and best practices to guide the creation and use of structured data elements and to continuously refine the many inconsistent and duplicate data elements that exist in the system. These inconsistent and duplicate data elements may have been caused by some of the same reasons described in this paper, such as predefined data element structure, the functionality of the editor tool, and constraints for reuse of data elements. We also acknowledge the rapid pace and size of EHR configuration and implementation projects. Our governance infrastructure will continue to remediate and refine structured data elements by focusing on clinical topics that apply across system modules and we have published elsewhere on these processes.

**Discussion**

Structured data elements are important for documenting the care of the patient and for collecting, storing and processing patient care information. In this study, we found that when we develop or implement different EHR systems (either homegrown or commercial products), we encounter various (similar and different) issues and challenges in building clinical data elements and relevant documentation assets used across different clinical settings. In this paper, we mainly used flowsheet data elements as an example, but one can image there are much more data elements needed for building templates and forms to compose clinical documents. Compared to some other coded data in EHRs (e.g., medications or diagnoses), they tend to be more like a “natural” language that contains rich, detailed information with various expressions. Often, they are difficult to be structuralized or standardized across different clinical settings and specialties. Without a comprehensive and efficient strategy, the data elements will grow rapidly to in order cover the diversity.

We faced many challenges to efficiently create and manage clinical data elements in a consistent, reusable, and interoperable fashion. Common themes, requirements, and desirable characteristics for controlled medical
terminologies have been described in previous studies (e.g., Cimino’s Desiderata\textsuperscript{15, 16}). Goossen et al provided a review of related work in the area of detailed clinical models.\textsuperscript{17} Oniki described lessons learned in detailed clinical modeling at Intermountain Healthcare.\textsuperscript{9} We have conducted a series of evaluations of the content created for our clinical systems at Partners Healthcare System and reflected on why the desirable characteristics outlined above have not been realized. Importantly, we consider our organization to be process-driven, motivated, and well-resourced; yet still struggle with these challenges outlined above raising the question: “How extensive is this problem at other organizations?” In the following, we summarize and discuss some key challenges and possible solutions that we have learned from our experience. We don’t intend to propose data element desiderata or comprehensive solutions in this paper. Instead, we start with basic but important challenges and practical solutions (i.e., challenges 1-3), and then move onto issues and solutions that are important and useful (i.e., challenges 4-6), but they may not achieve an expected impact and need to be combined with other approaches. Finally, we discuss more complicated challenges that need more extensive investigation and work in the future (i.e., challenges 7-8). It is hoped that these discussions may be useful in real-world practice for creating and maintaining clinical data elements.

1. Naming Convention, Attributes and Data Types

Well-defined clinical data elements are important to ensure data accuracy, accessibility, consistency and completeness. The institution should create a guideline for naming convention. When specifying the name for a data element, a preferred description may be chosen, for example, based on standard medical terminologies, and other lexical variations are used as synonyms. The design of data elements and attributes should promote reuse in different templates/forms and sharing among different practices. The value set should be generic (including common items) which allows users to choose the items that they need to create customized templates and forms.

EHR systems often have their inbuilt naming conventions and data types. Due the original design of the EHR system, these conventions may not follow the best industry practices. For example, they may tend to use composite terms for naming their data element (e.g., indicating site, specialty, or purpose within the name) and they may have their own predefined limited data types. Institutions, who implement such an EHR system, may consider developing principles and best practice for naming conventions, including how to name a data element, define its attributes and assign appropriate data types. For example, one may consider what components should be included in the name structure given a certain character limit in the naming field and in which order (e.g., if a data element is designed for a specific purpose and should not be used for other purpose, an indicator for an application or specialty may be added). Metadata may include unique identifier, owner, revision history, type, lifecycle state (e.g., draft, approved, retried), creator, date created, date last modified, sources, long name (i.e., descriptive name with more details about the content and intent), short name, reference terminology, etc.

2. Links to Standard Terminologies

Structured data elements need be encoded in computable forms to be used for clinical decision support, reporting, and other analytical tasks. Linking data elements to standard terminologies also has the potential to help identify duplicate, improve search and reuse. Although this is a basic and important strategy, it frequently gets overlooked. Multiple significant challenges for encoding standardized data elements exist. For example, Kim et al. found that SNOMED terminology provided both complete and partial matches of the Intensive Care Unit (ICU) nursing flowsheet data.\textsuperscript{18} The gaps in matching were felt to be mainly due to a lack of appropriate terms used in the original flowsheet and limitations of the concept models.

Manual review and mapping to standard terminologies is tedious and labor intensive. There are many lexical variations (including synonyms, local jargons, abbreviations, misspellings, and other specific symbols) and a lot of data elements have nested structures, thus exact string match or simple algorithms may be inadequate, indicating a need for semi-automated approaches and an efficient tool for mapping (such as the use of natural language processing). Since data elements for structured documentation contains concepts from multiple clinical domains, we need either adopt a comprehensive medical terminology or adopt and combine multiple terminologies at the same time. As mentioned above, similar clinical concept can be defined in both structured data (e.g., problem list) and as a data element for structured documents. This might need to define a scope, i.e., what types of concepts can be created as data elements and what concepts should be reused from other sources. For certain data elements (e.g., flowsheet and health maintenance items), one possible solution might be to keep these separated set of data elements and map them to standard terminologies. For these that are created separately in multiple sources, mapping them to standard terminology may facilitate future data integration and analysis.

3. Versioning and Change Management

Content lifecycle management is a basic and very critical problem. Versioning and change management is important in following the lifecycle of a specific data element. Once a data element is modified (e.g., add an
additional attribute), if the change management is inappropriate, this change may be propagated to every template and structured document using that data element and may potentially break these documents. This requires the system to track state changes (e.g., draft, in review, active, retired) and versions of these knowledge objects. In addition, change management also provides insight for future efforts to see why a particular element was created, the reasoning and thought process into why it was changed, and may provide guidance on when new but contextually related data elements are created. Most editors built within an EHR system often do not have a sophisticated capability for versioning and change management; therefore, additional content lifecycle management and an efficient edit/management tool outside an EHR may be needed.

4. Data Governance

Healthcare institutions should establish a set of principles and processes to ensure that clinical data elements and templates are formally managed throughout the enterprise based on applicable hospital accreditation standards, federal, state and institutional regulations, payer requirements and professional practice standards. The authority and accountability of data and knowledge assets should be clearly defined and enforced. Relevant parties within the organization should reach and follow an agreed-upon model, policies, best practices and guidelines which describe, for example, who can take what actions with what data elements, and when (under what circumstances) and how (using what methods and going through what processes). A committee or team that consists of executive leadership, clinical domain experts, informaticians, application analysts, and data stewards should be allocated, who employs certain methodologies and tools for managing, monitoring and improving data elements across the enterprise. For example, the committee may periodically conduct quantitative and qualitative analyses to check the quality and consistency of existing data elements and gather feedback from application teams and end users, and when necessary, refine data governance policies and procedures. Our PeC project has well-defined data governance strategies, but it has not eliminated all the problems, indicating it should be combined with other strategies such like those approaches mentioned above.

5. Monitoring, Communication and Training

We found that continuous monitoring of the quality and quantity of clinical data elements is needed to make sure that they are useful and reusable, and consistently meet users’ needs. We also found that the existing data elements in our legacy system (although some of them are duplicate and inconsistent) demonstrate the diversity of users’ needs. This resource demonstrating the diversity of users’ needs is now a valuable resource for us to use to create higher quality data elements. For example, we were able to analyze usage rates and aggregate multiple duplicate or similar data elements to create reference data elements.

Based on our previous experience, we found that when the users or analysts cannot easily find a data element that exactly meets their workflow, they tended to create or request a new one. Possible reasons for creating a new (duplicate) data elements include, insufficient time to look up existing content, insufficient tools to look up existing content, or insufficient training to look up existing content and meet best practices for data definitions. Effective communication between different stakeholders and training for data stewards and end users is critical to help maximize data element adoption among clinician users, reduce duplicates and facilitate linking to the standards.

6. Documentation vs. Structured Data Capture and Customization vs. Standardization

Each practice or provider may have their unique requirements or preference when creating a same or similar data element. Data elements should preserve details to some extent in order to capture something unique about the patient, the health problem, the practice, and the doctor. However, if each team works only on their own goals and ignore the needs of others, their silo requirements produces redundant information and functions and make integration with other silos very expensive, if not impossible. In contrast, standardization is critical for data reuse, integration and sharing. It is also important for high quality data capture and collection, and has impact on downstream data dependencies, including reporting, regulatory requirements, and CDS.

To alleviate the tension between documentation (e.g., note generation) versus structured data capture, it is important to build data elements in a more uniform form under a well-defined infrastructure. Development of individual information technology solutions at each practice or provider level with little coordination at the corporate level should be avoided. An enterprise-wide strategy may be established to ensure deeper cooperation and formal coordination among different parties (such as clinicians, application analysts and data stewards). For example, instead of allowing each local application team to create their data element directly into the system, a centralized vetting and management process should be in place. A possible workflow may be that the application team at each site send request to create a new data element to the enterprise committee. The committee will review whether or not a same data element already exists. If not, they will check whether the proposed data element follows the best practices. A management methodology is needed for standardizing the clinical data elements, while leaving
room for flexibility and customization. The effort required to implement this workflow during an EHR configuration is large and requires early engagement, prior to any clinician content validation efforts.

7. Pre-coordination vs. Post-coordination

As mentioned by Oniki et al.,

modeling and creating clinical data elements is not an exact science. A data element builder may face multiple choices when creating a data element, particularly with those that have multiple components. There are pros and cons of each approach and no state-of-the-art rules for deciding when to pre-coordinate and when to post-coordinate. Basically, post-coordination is more structured, so it is easy for computer to process and use for tasks such as information retrieval and clinical decision support. On the other hand, pre-coordinated terms are more like natural language and therefore more user-friendly. Oniki et al has proposed a set of general principles for helping make the decision when to pre-coordinate and when to post-coordinate. Such a decision is often constrained by the functionality that an EHR system provides and the burden created for end-users. For these cases, well-defined naming conventions may be helpful to create useful and consistent clinical data elements.

8. Reference Information Models

There are some existing efforts for structured clinical documentations as mentioned above, such as HL7 Clinical Information Modeling Initiatives, OpenEHR Archetype Model, Intermountain’s Detailed Clinical Models, etc., but there are also challenges and gaps in using these standards in real-world practice. Multiple factors affect the standard adoption, mostly due to the complexity of these models and the limited functionality of the EHR system and informatics expertise within local institutions to implement these models.

One of our endeavors is to create enterprise-wide reference data element models based on usage statistics. That is, we identify highly used clinical topics and convene an inter-professional panel of subject matter experts to vote and validate a reference model for that topic. Based on these activities, we create enterprise-wide reference data element models (e.g., with a comprehensive set of attributes) that specialty areas can customize based on their need (e.g., by selecting specific attributes needed for their purposes).

Summary

As discussed in Lean methodology,

we cannot only rely on clinicians, practice staff, and site analysts that work on a system to adhere with principles from training and generic advice to not create duplicate or inconsistent data elements. Individual institutions need to develop a comprehensive strategy by considering multiple factors. By designing and deploying a documentation system that fits the workflow of clinicians, it may have the potential to save provider’s time for documentation and eventually improve the quality of patient care.

Limitations of This Study

Our study has some limitations. This study was conducted in one integrated healthcare system. Our experiences may not be the same as what others have in other institutions. Our lessons were learned from developing and implementing particular EHR systems.

Conclusion

We presented three case studies of projects within Partners Healthcare to demonstrate the paramount challenges and issues regarding the development and management of data elements for structured clinical documentation in EHR systems. Some of the key challenges faced at Partners were discussed, and we offer some potential solutions to these challenges. It is hoped that our lessons and suggestions will be helpful and useful for institutions who face similar problems when developing strategies for managing their data elements and relevant clinical content.

Acknowledgments

This project was partially funded by the Partners Siemens Research Council. We thank Partners Clinical Informatics team, Acute Care Documentation team, Longitudinal Medical Records and the Partners eCare Structured Data Element Workgroup team for their work and support. We also thank Frank Chang for his suggestions and inputs for this study.

References

Sharing Patient-Generated Data in Clinical Practices: An Interview Study

Haining Zhu, BS1, Joanna Colgan, BA1, Madhu Reddy, PhD2,
Eun Kyoung Choe, PhD1

1Pennsylvania State University, University Park, PA; 2Northwestern University, Evanston, IL

Abstract

Patients are tracking and generating an increasingly large volume of personal health data outside the clinic due to an explosion of wearable sensing and mobile health (mHealth) apps. The potential usefulness of these data is enormous as they can provide good measures of everyday behavior and lifestyle. However, how we can fully leverage patient-generated data (PGD) and integrate them into clinical practice is less clear. In this interview study, we aim to understand how patients and clinicians currently share patient-generated data in clinical care practice. From the study, we identified technical, social, and organizational challenges in sharing and fully leveraging patient-generated data in clinical practices. Our findings can provide researchers potential avenues for enablers and barriers in sharing patient-generated data in clinical settings.

1. Introduction

Novel technologies have started to provide patients the ability to track their own health data1. Apple’s HealthKit and Google’s Google Fit are just two recently available platforms designed for capturing, storing, and retrieving personal health and fitness data. Startup companies, such as Fitbit2, 3 and Jawbone4, are also offering self-tracking devices. Self-tracking (or self-monitoring) has an extensive history in research and is noted to generally serve either therapeutic or assessment purposes, with assessments relying heavily upon data accuracy5. We build upon this preexisting framework for self-tracking and aim to understand how best to leverage Patient-Generated Data. In this paper, we define Patient-Generated Data (PGD) as health-related data created, recorded, or gathered by patients (or by family members or other caregivers) to help address their health concerns. These data include physical activity, sleep, food, and blood glucose levels. They can be collected using manual journaling, consumer health tracking devices, smartphone apps, or medical devices (e.g., blood glucose meter). In addition, PGD includes any independent lab results or tests ordered by the patients themselves (e.g. ubiome6, 23andMe7, and LabCorp8).

Enabling people to track their own personal health data provides the potential to leverage the data in medical contexts, but to do so requires an effective means of data sharing among patients and clinicians. Moreover, patients who suffer from a chronic disease often find that self-tracking is burdensome and the sharing of data adds to that burden9. Even motivated patients can be hindered by awareness of the disease, the scattered nature of self-tracking tools, and data collection burden, making PGD hard to collect, and thus share10. With the goal of designing technology that allows for effective and meaningful sharing of PGD among patients and clinicians, we conducted semi-structured interviews with both patient and clinician participants to understand how they currently share PGD and to identify what barriers they face. Based on our findings, we suggest design considerations for creating technologies to support PGD data sharing among patients and clinicians.

2. Background

In this section, we provide a background of the Quantified Self movement, patient-generated data, and patient-generated data sharing.

2.1. The Quantified Self Movement

Advancements in mobile and wearable sensing technologies have spawned a movement called the “Quantified Self.”11 Those who practice Quantified Self (Q-Selfers) track data (e.g., food consumption, mood, physical activity) by manual journaling or self-tracking technology for various reasons such as health behavior change and curiosity10. Coupled with participatory health initiatives, big health data, and trends in patient engagement, data provided by Q-Selfers will hopefully be freely exchanged across healthcare sectors and result in personalized preventive medical practices12. The shift to personalized preventive medicine might combat contemporary health challenges, including sleep disorders, mental health issues, and chronic disease management12. Recognizing these potential benefits, we note a growing number of patients who practice Quantified Self as an approach to preventive health or chronic health management10. 13. Barriers and enablers in the Quantified Self movement vary by individual and situation. Q-Selfers might find the inability to customize and add specific tracking parameters to a ready-made tool to be a
barrier; furthermore, ready-made systems that are intended for diagnostic purposes benefit from precision and accuracy to enable reliable diagnoses, while interactive systems require methods of collaboration and communication to enable the relay of contextually relevant data. Despite the increasing number of online services and available platforms (e.g., Electronic Health Records (EHRs), Personal Health Record (PHRs), electronic patient portals), adoption rates of Quantified Self in medical practices are low and the efficacy of these platforms’ abilities to empower Q-Selfers or improve health outcomes remain untested. Thus, we want to understand the barriers patients and clinicians face with the Quantified Self approach and devise ways to leverage the extensive data patients are generating.

2.2. Patient-Generated Data (PGD)

PGD have the potential to lead to novel medical insights and can promote positive changes in contemporary medical practices. For example, data generated by patients using continuous glucose monitoring systems and self-reported surveys have provided insights into complications in hypoglycemia patients, which might have been missed in verbal communication during clinical visits. PGD such as patient status and history have also been successfully utilized in clinical policy decision making, but this decision making would likely improve with the addition of new data categories, such as patient drug combinations, vitals, and blood results. Despite these potential benefits, however, there are inherent limitations to PGD, including the reliability and accuracy of data, forgetfulness of patients, innate attitudes towards technology, and patients’ self-bias. To utilize PGD in medical practices and research, these barriers need to be overcome. One way researchers have addressed these limitations is by suggesting ways to lower the patients’ manual tracking burdens and increase their adherence to tracking. To further lower the tracking burden, automated sensing is commonly embedded in self-tracking technologies (e.g., Lullaby, MyBehavior). In addition, systems that allow users to edit or add to data inferred by devices are more reliable and credible. Thus, innovative technologies and cultural changes related to the Quantified Self movement have allowed for patients to collect more accurate and reliable health data. These changes related to patient self-management and empowerment through tracking are therefore necessitating the creation of successful methods of leveraging PGD in medical applications.

2.3. Patient-Generated Data Sharing

Sharing can occur in many ways, including between a clinician and a patient and amongst patients, with each having unique challenges. Our work focuses on the contexts of sharing PGD between a clinician and a patient. From the patients’ perspective, self-tracking and data sharing can impose extra burden on top of other efforts for managing chronic conditions. Clinicians also have a less than positive view of sharing PGD; they have concerns about large amounts of PGD interfering with their abilities to provide timely and efficient healthcare to their patients. Sharing PGD can also impose privacy burdens on hospitals. Modern medical practices protect patient records with strict laws such as HIPAA. Although PGD is not traditional medical data, it has certain characteristics of it, and thus, there is a growing concern over whether it is subject to HIPAA when the data gets shared with clinicians via mobile devices. However, clear legal standards do not exist. PGD allows patients to actively engage in clinician-patient collaboration; such data can be shared during clinical visits (i.e., synchronous communication) or outside of clinical visits (i.e., asynchronous communication). Asynchronous communication tools, such as patient portals, create secure opportunities for dialogue, but this interaction often loses the benefit of non-verbal communication (e.g., using body language and tone); these tools are most beneficial if they compensate for nonverbal deficiencies and if used by motivated, technologically savvy patients. However, real-time tracking tools (i.e., mobile phone), which can be used for synchronous communication (i.e., in person, telemedicine, phone call), allow for more consistent data recording in patients. These tools might thus mitigate issues of patient compliance, but they often lack support for data sharing. Our work specifically aims to understand what impedes PGD sharing during clinical practices, including technical, social, and organizational challenges.

3. Method

We conducted semi-structured interviews with 21 participants (12 patient participants and 9 clinician participants). We wanted to identify the enablers and barriers inherent to sharing PGD for patient-clinician communication, and to gain insights into design requirements for future technology interventions. The purposes of this study are to help patients become active observers and participants in their care and to inform clinicians of PGD data to enable them to provide better care.
Participants: Participants were recruited by word-of-mouth referrals and advertisements in handouts placed in public locations, hospitals, and online (e.g., university mailing lists). In this qualitative study, we sought diverse perspectives from broad participants. To be eligible as a patient participant, individuals had to (1) have experience collecting PGD (e.g., sleep, diet, exercise) for themselves or for their loved ones (e.g., a parent tracking health data for his or her child); and (2) have experience sharing the collected data with clinicians. Patient participants were not limited to current patients. Patient participants did not require a clear diagnosis to be eligible. We excluded individuals who were minors or non-English speakers. Clinician participants were individuals who (1) regularly consulted with patients (e.g. therapist, physician); and (2) had experience seeing patients who wanted to share their PGD during the visit. Table 1 and Table 2 show our participants’ demographic information.

<table>
<thead>
<tr>
<th>ID</th>
<th>Background</th>
<th>Age</th>
<th>Gender</th>
<th>Years of practice</th>
<th>Tracking Initiation* (CI/PI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Physical therapist</td>
<td>31-50</td>
<td>M</td>
<td>17</td>
<td>CI/PI</td>
</tr>
<tr>
<td>C2</td>
<td>Internist</td>
<td>31-50</td>
<td>F</td>
<td>11</td>
<td>CI/PI</td>
</tr>
<tr>
<td>C3</td>
<td>Internist</td>
<td>31-50</td>
<td>F</td>
<td>6</td>
<td>CI/PI</td>
</tr>
<tr>
<td>C4</td>
<td>Internist</td>
<td>31-50</td>
<td>F</td>
<td>8</td>
<td>CI</td>
</tr>
<tr>
<td>C5</td>
<td>Primary care</td>
<td>31-50</td>
<td>M</td>
<td>24</td>
<td>CI/PI</td>
</tr>
<tr>
<td>C6</td>
<td>Psychologist</td>
<td>31-50</td>
<td>M</td>
<td>10</td>
<td>CI</td>
</tr>
<tr>
<td>C7</td>
<td>Pediatric nephrologist</td>
<td>31-50</td>
<td>M</td>
<td>10</td>
<td>CI/PI</td>
</tr>
<tr>
<td>C8</td>
<td>Internist</td>
<td>18-30</td>
<td>F</td>
<td>1</td>
<td>CI/PI</td>
</tr>
<tr>
<td>C9</td>
<td>Pediatric nephrologist</td>
<td>31-50</td>
<td>F</td>
<td>3</td>
<td>CI</td>
</tr>
</tbody>
</table>

*CI denotes tracking was initiated by clinicians (clinician-initiated), whereas PI denotes tracking was initiated by patients (patient-initiated).

Study Procedure: We conducted semi-structured interviews. Interviews were conducted either via phone (n = 10), Skype (n = 7), or in-person (n = 4). We also collected screenshots of how PGD was shared, including tracking tools, diary sheets or notes, and visualizations. We designed the interview to guide participants to discuss their firsthand experiences with self-tracking. Interview questions for patient participants included the following: (1) What motivated you to track and share data? (2) How and with whom did you share PGD? (3) What were barriers and enablers to sharing PGD? (4) How did the activity of tracking and the knowledge it provided affect your general life? Interview questions for clinician participants included the following: (1) How do patients typically share PGD? (2) What are some barriers that prevent sharing PGD during clinical visits? (3) How could PGD be utilized to support the clinician and care team? From these interviews, we gained a deep understanding of the key barriers, which inhibit clinical utilization of PGD and ways in which we might mitigate the identified barriers. The research was approved by the Institutional Review Board.

Data Collection & Analysis: All interviews were audio-recorded and transcribed to facilitate analysis. Each interview ranged from 30 to 90 minutes. To ensure the confidentiality of participants, we assigned a unique participant identifier to refer the roles of the participant: C# to denote a clinician participant, and P# to denote a patient participant. Two researchers analyzed each transcript of the individual interviews and used qualitative open coding to note prominent themes that were discovered across the data pool. Once the team agreed on the high-level themes identified, one researcher then iteratively coded the data and updated the coding scheme. The high level

<table>
<thead>
<tr>
<th>ID</th>
<th>Medical / Health Condition</th>
<th>What they track</th>
<th>Current patient (Y/N)</th>
<th>Age</th>
<th>Gender</th>
<th>Tracking initiation* (CI/PI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Pre-hypertension</td>
<td>Biking, cholesterol, blood pressure, blood sugar</td>
<td>Y</td>
<td>51-65</td>
<td>M</td>
<td>PI</td>
</tr>
<tr>
<td>P2</td>
<td>Pregnancy</td>
<td>Pregnancy symptoms, activity and calorie tracker</td>
<td>Y</td>
<td>18-30</td>
<td>F</td>
<td>PI</td>
</tr>
<tr>
<td>P3</td>
<td>Family history of cancer</td>
<td>Blood tests, telomere testing, neurofeedback, fecal biome test</td>
<td>Y</td>
<td>&gt;65</td>
<td>M</td>
<td>PI</td>
</tr>
<tr>
<td>P4</td>
<td>Coronary artery disease</td>
<td>Weight, exercise, training, steps, lab results, key life events</td>
<td>Y</td>
<td>&gt;65</td>
<td>M</td>
<td>PI</td>
</tr>
<tr>
<td>P5</td>
<td>Chronic sleep syndrome</td>
<td>Exercise, sleep, period, weight, heart rate</td>
<td>Y</td>
<td>31-50</td>
<td>F</td>
<td>PI</td>
</tr>
<tr>
<td>P6</td>
<td>High-level of blood glucose</td>
<td>Blood pressure, blood glucose, annual labs</td>
<td>N</td>
<td>51-65</td>
<td>F</td>
<td>PI</td>
</tr>
<tr>
<td>P7</td>
<td>Restless leg syndrome</td>
<td>Weight, physical activities, blood pressure, heart rate, food journaling</td>
<td>Y</td>
<td>51-65</td>
<td>M</td>
<td>PI</td>
</tr>
<tr>
<td>P8</td>
<td>Mycosis fungoides &amp; lung cancer</td>
<td>Medications and medication reactions</td>
<td>Y</td>
<td>51-65</td>
<td>M</td>
<td>PI</td>
</tr>
<tr>
<td>P9</td>
<td>Artificial heart valve replacement</td>
<td>Cardiovascular fitness, steps, cholesterol, triglycerides, sleep, weight</td>
<td>Y</td>
<td>51-65</td>
<td>M</td>
<td>CI</td>
</tr>
<tr>
<td>P10</td>
<td>Type I diabetes</td>
<td>Blood sugar, weight, carb counting</td>
<td>N</td>
<td>18-30</td>
<td>M</td>
<td>PI</td>
</tr>
<tr>
<td>P11</td>
<td>Autoimmune conditions</td>
<td>Symptoms, activity, humidity, medication, stress level, sleep, period</td>
<td>Y</td>
<td>31-50</td>
<td>F</td>
<td>PI</td>
</tr>
<tr>
<td>P12</td>
<td>Chest muscle problem</td>
<td>Ride, time, speed, distance, and heart rate</td>
<td>N</td>
<td>18-30</td>
<td>M</td>
<td>PI</td>
</tr>
</tbody>
</table>

*CI denotes tracking was initiated by clinicians (clinician-initiated), whereas PI denotes tracking was initiated by patients (patient-initiated).
categories we include are as follows: motivations for patients to collect and share health data; motivations for clinicians to adopt PGD; methods and means to share PGD; tensions between clinicians and patients; barriers and drawbacks to PGD sharing; and enablers and workarounds to share PGD.

4. Results

In this section, we present the results from the interview study to report existing gaps in PGD sharing. In particular, we describe the nature of patients’ tracking, address the difficulty in sharing data between patients and clinicians, and report the barriers which inhibit the utilization of PGD—from both patients’ and clinicians’ perspectives.

4.1. Clinician-Initiated versus Patient-Initiated Tracking

In the clinical context, tracking was initiated either by a patient or clinician. In our patient dataset, most of the patients initiated the tracking without any prompting from clinicians (Table 2, last column). On the other hand, all clinicians we recruited had experience initiating the tracking by providing patients with a diary and asking them to track certain items (Table 1). Moreover, most of the clinician participants had experienced patients bringing their data to visits and sharing it voluntarily. We observed that clinicians’ receptiveness to PGD varied depending on who initiated the tracking. When clinicians initiated the tracking, it was often the case that they required PGD for a specific medical reasoning, thereby the value of the data was high. In these instances, patients became a “diagnostic agent” for clinicians, and thus played an active role in personal diagnoses. Patient-initiated tracking, however, was not always welcome by clinicians, which we further describe in the following.

Clinician-Initiated Tracking: Clinicians asked patients to self-track for various reasons. C3 was particularly interested in pushing patients for engagement purposes. C3 said, “…knowing the actual data is less important than just hearing their perspective and knowing where they’ve come from and engaging them on that... if they’re tracking... addressing it is a way to sort of reinforce their engagement.” Clinicians’ authority helped with motivating patients to track, as C5 perceived: “it’s very rare that they [patients] record anything unless you ask them to.” Most of the clinician participants noted it was common to ask patients to provide data such as reactions, side effects, and symptoms when they receive new medications. The practice of asking for data and making decisions based on it was essentially what it meant to be a clinician. Clinician participant C5 suggested that clinicians initiated tracking for the purpose of accurately assessing a patient’s issue: “Maybe for sleep disorders and sleep insomnias, when people complain to that, you try to get them to do a two-week sleep log so you’ll really know how they’re sleeping.” Tracking was most often initiated by clinicians in cases with chronic diseases (e.g., diabetes, hypertension) or persistent symptoms: “Once you diagnose a patient with diabetes, you’ll send them to the diabetic educators, we’ll teach them how to use a glucometer and then go over with them kind of how to keep track of their blood sugar values…” [C8]. Participant P9 also explained that clinicians initiated a tracking regimen after certain medical procedures or during rehabilitation: “After I had heart surgery, I was in a cardiac rehab program for several months and that was a daily or a three times/week tracking.” Finally, C8 explained that PGD is useful for discussing lifestyle changes and barriers preventing patients from achieving their health goals.

Patient-Initiated Tracking: The most common purpose for a patient to initiate self-tracking was to develop self-awareness and self-management skills, with 8 participants offering similarly themed explanations. Patient P11 stated, “…if I’m going to go out and drink… I will expect to have a bad night of sleep. So if I need to be ready to give like a presentation, I’ll just plan for that and not go out. I feel like I know better how things affect me and how long they’re going to affect me. It has helped me optimize myself.” In cases with patient-initiated tracking, another common tracking motivation was individual curiosity—for example, curiosity about how one’s lifestyle affects weight, or whether a treatment works. P8 remarked, “Gathering that information or looking at that information has given me hope…This drug that I have is effective at first but doesn’t seem to be as effective right now… we may discontinue it and start something else.” Patients with medical problems, such as diabetics, often tracked data related to those medical problems. Three patients tracked data so that they could better collaborate with clinicians or answer their questions more effectively. Two other patients stated that they shared data with clinicians to initiate conversations and keep them informed. P9 specifically noted, “I think tracking helps me feel like I am participating in my health and my medical care directly and that I’m being proactive about it. It enables me to have informed conversations with my medical providers that are based on data.” Furthermore, some patients wanted to show the data to collaboratively alter their medications, or to come off medications. P4 noted “…it’s very easy to say, ‘Your cholesterol is too high, take more medicine because we know you’ll go down 10% if you double the dose.’ But by tracking I did, I’m able to say, ‘Hey, if I get my weight down 12 pounds, I will reduce my cholesterol 20 points and do the same thing that way.’” Five patients mentioned this desire to collect data to help change treatment plans. As such, many patients appeared very eager and motivated to be equal partners in the promotion of their health, and the
fact that these patients track multiple factors suggests that some are trying to promote whole health, rather than treat specific symptoms. Lastly, a clinician participant mentioned that patients’ tracking can potentially be due to the patients’ mental health disorder or obsessive behavior.

4.2. How is PGD Shared

Patients track a wide array of health information. In our study, patients tracked factors such as exercise, treatment changes or medically relevant factors, food, alcohol, caffeine consumption, travel and lifestyle, symptoms and ailments, weight, period data, sleep and sleep patterns, test results, general health history, and chronic or persistent health problems (e.g., diabetes, high blood pressure). We examined how data is shared among patients and clinicians by employing the space-time matrix. According to the space-time matrix, PGD sharing can be categorized based on where it occurs (distributed vs. co-located) and when it occurs (synchronous vs. asynchronous) (Table 3).

Table 3. How Patient-generated Data is Shared: Space-time Matrix.

<table>
<thead>
<tr>
<th>Co-located</th>
<th>Distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synchronous</td>
<td>During the visit (i.e., in-person)</td>
</tr>
<tr>
<td>Asynchronous</td>
<td>-</td>
</tr>
</tbody>
</table>

Synchronous Co-located Sharing: Verbal communication during visits was the most common data sharing method in clinical practices. Most participants had shared data verbally during visits, as C5 reported: “[Patient typically shares data during the visit] because that’s the only time we have contact.” Some patients collected their personal data but chose not to share it in any other form but verbally so that they could maximize face to face time with their clinicians: “You get a ton of data, but if you give him all these numbers and stuff to go through, then you’re going to bleed out the time for the visit…” [P4]. There were two cases in which a patient specifically mentioned why they did not use technology to share data with a clinician. Those two instances were specifically the inconvenience of traveling with a laptop and the uncertainty of proper etiquette. P12 noted, “I was worried that he [clinician] might feel annoyed if I pulled out my mobile phone… Normally clinicians don’t want patients to pull it out.” Using a tool (e.g., spreadsheets, images, timelines, videos, data visualizations) was the second most common way of sharing data during visits; the most common tool used by patients was paper. C3 confirmed this by stating, “They may just have a piece of paper with everything they ate for the last 48 hours. Or they might have an ache diary…. It may just be written on a piece of paper.” Information visualizations were thought to be good tools for quickly sharing data to uncover significant insights, especially for those who had the skillset to create visualizations. For example, P11 was a designer who could draw a timeline using Adobe Illustrator, which she printed out and shared with clinicians. P11 expressed, “I think the clinicians have appreciated the timeline format… a visual timeline like a graph, a timeline on a graph would be helpful for sharing (Figure 1).” Patients commonly used data collection tools, such as medical devices and wearable trackers, as data sharing tools, because there were few available tools explicitly designed for data sharing. C8 noted, “Some patients who feel like it’s going to take too much time and efforts for them within their lifestyle will just bring in the glucometer because it’s easier for them to just bring it in.” Patients also presented clinicians data on their mobile phones or desktop applications. Some patients even brought a computer to a clinical visit to demonstrate trends in data. Although any device with the ability to store and present data could potentially be utilized for PGD sharing, clinicians found it especially difficult to manage the data presented by data collection tools. As stated by C8, “…trying to kind of figure out how to maneuver and like data summary of the data from each different app, can be a little challenging and time consuming in an acute visit.”

Figure 1. P11 used Adobe Illustrator to create a medical timeline, and brought the printed version when visiting to a new doctor. The peach color represents gastrointestinal symptoms, and the blue represents Myasthenia Gravis.
Synchronous Distributed Sharing: Phone calls offered a quick and convenient way to share data, especially when other electronic means such as emails could not be utilized or when patients wanted immediate responses but were unable to make an appointment. Furthermore, phone calls provided a benefit over email communications, as they provided important cues such as tone of voice as well as opportunities for clinicians to ask for clarification. The majority of clinician participants asked patients to give a phone call in certain situations: “I get most of my data from the patient during their actual visit…. there will be times when I will ask them like if I make a change in their medication regimen to keep track of their sugars for the following week and call the clinic with those numbers so I can make further adjustments over the phone if necessary” [C8]. Telemedicine offered unique opportunities and drawbacks, and clinician participants had conflicting opinions towards it. Two clinician participants agreed that telemedicine could help them see more patients, and that more patients would benefit from their treatments if clinicians could provide electronic or telephone based communication. Two other clinician participants however questioned the accuracy and reliability of the data shared via telemedicine tools (e.g., phone): “If they’re giving you information over the phone, can you trust that their sugar levels are recordable?... it’s an electronic device that doesn’t transmit it electronically.” Other participants had neutral reactions toward telemedicine but potentially showed a slight preference for email communications. Currently, synchronous distributed sharing is typically uncompensated effort, however, and this significant drawback might have resulted in clinicians being more reluctant to look at PGD and more resistant to using technology outside of clinical visits. C2 explained, “...it’s better that it [PGD] comes to me in an office visit…. I do a lot through like on patient portal and do a lot of management that way but we have no way of compensating that time…. it’s a lot of uncompensated effort on my part… that’s something that I do it because I wanted to be I want them to get better.” Although the work was uncompensated, some clinicians still asked patients to send data via email or phone, took time to look at the data, and offered medical opinions because they believed that it is beneficial to patients.

Asynchronous Distributed Sharing: Asynchronous distributed communication was beneficial in situations where patients and clinicians had conflicts of schedules. Faxes, emails, or patient portals could enable the sharing of data immediately, while allowing for a clinician to respond when he or she has the time to do so. C8 noted, “...via the patient portal, they can actually send a message directly to me if they have any questions or concerns between visits.” Patient P4, who has a software engineering background and is a retired medical doctor, was the most extreme and unique case. He designed and built his own software, which assisted in the sharing of data in an asynchronous distributed manner (Figure 2-c, 2-d): “I just put him [clinician] to my websites... if he just puts in my URL that he can access it from the note, and it’s actually easier for him.”

Figure 2: Example of screenshots that patients shared. (From left to right: P1, P11, P4, P4). P1 and P11 used excel spreadsheets to track their health data, here shown as (a) and (b). P4 created software that generates graphs to show the correlations of certain fitness data, such as Fitbit steps, Aerobic time, and weight.

4.3. Tensions around PGD

Tensions existed between clinicians and patients, which discouraged the sharing of data. Not all clinicians were receptive to patient-generated data. We report major tensions existing in the current medical practice.

Tensions Between Clinicians and Patients: The first tension we observed was the misalignment of the clinician’s agenda and the patient’s expectations; this includes such things as patients recording clinically irrelevant data, or not recording relevant data, and differences of opinion between clinician and patient regarding healthcare needs. C2 noted, “Because sometimes their ideas of why they’re there are different from my ideas of why they need to be there. Then, we have a discussion.” We also observed that patients’ desires to maximize their time with clinicians often created tension between them and clinicians. In many of those cases, clinicians were unable to provide time outside of their visits. One clinician participant expressed, “you don’t have time outside of that [in-person visit]. In
medicine, it has to be as the patient is visiting. Because otherwise, you’re taking time either from somebody else or from yourself” [C5]. Many clinician participants and a few patient participants were concerned that clinicians could be overwhelmed by raw PGD. Many participants specifically expressed concerns about clinicians receiving too much data at once and not knowing how to process it. “If people come in with so much information that you can’t make sense of it, it doesn’t help clarify things...” [C3]. The third tension, experienced by most clinicians, was the issue of the reliability and accuracy of PGD. “It’s not usually accurate because they don’t have all the data to support it so they’re kind of giving you this memory of what they think” [C4]. Concerns regarding data reliability and accuracy were not limited to patients’ data reporting errors. Clinicians also pointed out that uncalibrated machines, improper testing protocols, transcription errors, device errors, or some other unknown factors could result in unreliable or inaccurate data. Thus, clinicians were mistrustful of PGD shared by their patients unless accuracy can be verified. C8 noted, “...people just don’t actually check their blood sugars routinely enough, and if they do they kind of do it sporadically and don’t write down the exact date and time, etc. (...) that can be a little bit more challenging.” Although data collected by a patient could be entirely accurate, and thus reliable, it was not meaningful to a clinician if it did not contain all of the data that the clinician might need.

**Attitude towards using technology:** Clinicians responded differently and sometimes ambivalently to the use of technology for tracking and sharing PGD. About half of the clinician participants expressed preference for paper based data tracking over technology-assisted tracking. C3 noted, “...The electronic stuff is a lot easier to collect, but I think it’s a lot harder to analyze because it’s less common to be able to see everything all on one screen....” However, half of the clinician respondents also reacted favorably to the use of technology for tracking, or negatively to paper based tracking. These groups were not mutually exclusive; there were individuals who both preferred paper methods and reacted favorably to technology-assisted methods. A few clinician participants reacted neutrally towards technology. C3 suggested her reaction was pragmatic, and said, “So whatever method is easiest for them to get you good accurate data.... My patients would lose the piece of paper, if they did something on their phone they would do better with that.” Our data suggests that the clinician’s reaction to technology versus paper-based method is a multifaceted issue; clinicians’ responses to the use of technology likely corresponded to their initial attitudes towards technology in general. Clinicians tended to believe that both paper and technology-assisted method have their own merits and pitfalls as well as niches, which could explain the apparent ambivalence or even neutral responses noted by participants.

**Table 4:** Barriers and Enablers to Leveraging Patient Generated Data.

<table>
<thead>
<tr>
<th>Barriers</th>
<th>Supporter</th>
<th>Enablers</th>
<th>Supporter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inability to transfer PGD to electronic medical records</td>
<td>P4,P7,P9,P11,C1,C2,C3,C4,C7,C8,C9</td>
<td>Integrate PGD into clinicians' workflows</td>
<td>P4,C7</td>
</tr>
<tr>
<td>HIPAA intruding into medical care</td>
<td>C1,C5,C6,C7</td>
<td>An amicable sharing platform and automatic upload of PGD</td>
<td>P1,P7,P11,P12,C1,C4,C7,C8,C9</td>
</tr>
<tr>
<td>Data security and privacy concerns</td>
<td>P3,C1,C2,C3,C5,C6</td>
<td>Portable and shareable patient-stored medical records</td>
<td>P3,P4,P7,P9,C1</td>
</tr>
<tr>
<td>Difficulty with accessing and using PGD within EMR</td>
<td>P12,C1,C2,C5,C6</td>
<td>Using data visualization and summarization for efficient sharing</td>
<td>P2,P7,P9,C2,C4,C6,C7</td>
</tr>
<tr>
<td>Limited amount of clinicians' time during visits</td>
<td>C1,C3,C4,C5,C8</td>
<td>Sending data to clinicians prior to visits</td>
<td>P11,C2,C5,C7,C9</td>
</tr>
</tbody>
</table>

**4.4. Barriers and Enablers to Leveraging Patient Generated Data**

Most clinicians consistently noted the difficulty or inability of transferring PGD to EMR as a barrier for data sharing. C4 said, “…it [PGD] doesn’t go into the electronic health record. So unless your clinic is set up where a nurse practitioner or someone else can go through all of that data, set that up, make sure they have it and they give that information to the clinic, it’s not really helpful to anybody else.” A related concern was HIPAA and data security and privacy issues as C5 explained: “We have to send them, ‘We cannot respond to this [PGD], this is a violation of HIPAA. Just go away or go to a secured server’” [C5]. However, it was not clear whether EMR was the right place to store all the patient-generated data collected outside the hospital because (1) there is currently not a good user interface within EMR to upload and view PGD; and (2) the data can be quickly buried in the EMR system. In fact, P4 stated, “…since I have the data, if I get referred to a specialist, I'm not counting on his transfer... I wrote up a little summary with my graphs and charts... [the specialist] said, ‘I just read what you had, it was easier.’” We also learned that storing PGD in the EMR system was not just a data integration issue, but a bigger issue of integrating PGD into clinicians’ workflows, making the data easy to access, use, and transfer. A patient participant noted this challenge: “Patients are self-trapped, when they come in with data that’s difficult for the
doctor to digest and the doctor doesn't like that. (...) The key thing is, that's the work for the patient but it has to create this thing that's so easy for doctor, so that make their life easier too or they'll find a way to ignore it” [P4].

A majority of clinician participants mentioned time constraints as an important barrier impeding data sharing, which was apparent to patients as well. P9 noted, “I have to prioritize what I think is important to share. I’m sensitive that I don’t get a lot of time with them that they have a lot of information to process” [P9]. Participants shared ideas for efficient PGD sharing, for example, sending data to clinicians before visits, as noted by C2: “So if a patient can handle technology and can send the information in preparation for a visit...That would be great.” Likewise, data visualization was noted to have aided efficient data sharing. “I think the visualization is great... So we know there are high levels of innumeracy in patients so figuring out ways to present the data, not just for the patients, but me so I can get a quick overview...” [C4]. We believe that providing insights using visualizations in the context of PGD sharing opens up many possibilities for future research, especially designing a PGD sharing platform that can ensure quick and efficient sharing while providing insights.

5. Discussion

In the previous section, we highlighted a variety of issues related to the sharing of PGD between patients and clinicians. We now turn to four discussion points surrounding PGD sharing. These issues are (1) aligning clinician/patient perceptions about PGD by including clinicians in the early stages of patients’ tracking; (2) integrating PGD into clinical systems; (3) aligning organizational policies; and (4) empowering patients as partners in healthcare.

First, it is important to align the clinicians’ and patients’ perceptions about PGD as closely as possible. Both clinicians and patients are interested in seeing health trends and patterns or documenting treatment efficacy, but the two groups might have very different perceptions about the type and value of the data. One of the primary motivations for patients to collect PGD was to gain an as in-depth picture as possible of their health, which they can then share with their clinicians. Based on our interviews, however, we learned that when patients initiated tracking, clinicians often ignored the data collected and did not find the data useful, especially when it was either incomplete or irrelevant to the current issue. For clinicians to leverage the data, clinicians must know that data is accurate, reliable, and aligned with their agenda. We therefore see opportunities in making the PGD more useful and relevant by involving clinicians early on to provide guidance to patients on how and what data to collect and to help patients understand why such data is important. With such guidance from clinicians (e.g., what to track, how often to track, what format to use), patients might be able to collect more clinically relevant data, thereby making clinicians more likely to leverage that data in their clinical practice. Likewise, if patients could provide contextual information regarding their testing and data collection, clinicians might be more accepting of PGD.

Second, although PGD could have immense value for patient care, clinicians cannot effectively use this data until it is integrated into the clinical systems. The integration will require the development of a technical infrastructure to transfer the information from the patient to the clinician. Although patient portals have started to be used for sharing patient information with clinicians, it is not being used for PGD data sharing, especially the synchronous co-located manner, which was the most prevalent way of sharing PGD. Data visualizations, analytics, and summarization techniques will play an important role in making PGD valuable for clinicians and patients, helping them quickly gain critical insights and reflect on the data.

Third, clinicians need to be incentivized to incorporate PGD into their everyday workflows. While the technical challenges of integrating PGD into clinical systems are important, it will be just as important to re-align organizational policies to incentivize clinicians to utilize PGD. It is unclear at this point whether sharing PGD will lead to more patient encounter time (if sharing is done inefficiently) or even less patient encounter time (if sharing is done efficiently) than what is currently allocated for patient encounters for clinicians. The effectiveness of sharing would depend on how a PGD sharing platform is designed and how well the data is presented to the clinicians.

Lastly, clinicians need to understand and adapt to the cultural shift in healthcare, in which more patients are attempting to make healthcare a collaborative endeavor. The patient participants we talked to wished to alter their treatment plans and create conversations with their clinicians; they wanted to be equal partners in healthcare and careful observers of their bodies, as opposed to traditional passive patients. Furthermore, some of these patients promoted their whole health through employing tracking for preventive purposes, rather than to treat an illness, thus changing the current paradigm of receiving healthcare. Patients are also motivated by clinicians; they will engage in their healthcare more if clinicians welcome their involvement. One way of inviting patients into the care team is through encouraging patients to start tracking their data in a way that can help clinicians.
6. Limitations and Future Directions

Our study’s limitations include the small participant sample size and purposive sampling towards receptive patients and clinicians due to our inclusion criterion of requiring prior PGD collecting and sharing experience. The study results are therefore unlikely generalizable to the entire population. However, our goal was to discover current practices and barriers in patient-clinician data sharing, and we argue that our inclusion criteria are necessary to gain meaningful data, as we were looking for initial insights into existing phenomena. Our findings could be used to create an in-depth study of PGD sharing behaviors, by creating a prototype intervention and testing it with a representative sample. We also selected participants from multiple localities and medical subdivisions, and with various health conditions to recruit a diverse population, because we were interested in the coverage of various issues. We also note that the patient participants were not necessarily the clinician participants’ patients; this is why most of our patient participants indicated that they had initiated their tracking (patient-initiated, or PI), but all of our clinician participants indicated that they initiated the tracking (clinician-initiated, or CI) or both CI and PI. Therefore, our patient participants might have been particularly motivated in collecting and sharing PGD whereas our clinician participants might have actively encouraged patients to track. In the future, we are planning to study patient and clinician pairs, so that the selection bias may be reduced. The findings of this study will help us design a future platform for PGD sharing to enhance patient-clinician communication.

7. Conclusion

In this study, we described the current practices of PGD sharing among patients and clinicians. We particularly examined the differences between tracking initiated by clinicians versus patients. We then employed the space-time matrix to analyze various PGD sharing practices. After describing tensions between clinicians and patients, we identified barriers and enablers to PGD sharing. Many of the barriers were related to the lack of infrastructural support for data integration, but this was not merely a technological issue; a bigger challenge was how PGD sharing could be incorporated into clinicians’ workflow. Most of these barriers can compound to further inhibit clinician-patient communications. Understanding a clear picture of how PGD is currently shared and its associated challenges are important for the design of systems that can leverage PGD and integrate the data into clinical practices. This initial study provides new opportunities to study PGD sharing between patients and clinicians with better focus on the significant issues we identified. We are particularly interested in how information visualization can be designed to aid efficient data sharing. With the ever-increasing growth of personal health mobile applications and wearable sensors, patient generated data will only increase. We are only starting to understand the challenges that we need to address to integrate this data into the healthcare process.

8. Acknowledgments

The research presented in this article was funded by the National Science Foundation under awards CRII:CHS-1464382.

References

7. 23andMe. https://www.23andme.com
8. LabCorp. https://www.labcorp.com/wps/portal/services
15. PatientsLikeMe. https://www.patientslikeme.com/
27. Deven McGraw JD, Pfister HR, Ingargiola SR. Lessons from Project HealthDesign.
34. Osborn CY, Rosenbloom ST, Stenner SP, Anders S, Muse S, Johnson KB, Jirjis J, Jackson GP. MyHealthAtVanderbilt: policies and procedures governing patient portal functionality. JAMIA. 2011 Dec 1; 18(Supplement 1):i18–23.
Clinical Decision Support to Detect High Risk Patients with QTc Prolongation

Christopher A. Aakre, MD, J. Martijn Bos, MD, PhD, Gyorgy Simon, PhD, Robert F. Tarrell, Michael J. Ackerman, MD, PhD, Pedro J. Caraballo, MD
Mayo Clinic, Rochester, Minnesota

Abstract
Prolonged QTc is a risk factor for sudden cardiac death. To identify patient at risk, we implemented clinical decision support integrated in the electronic health record to detect new QTc prolongation of 60ms and alert clinicians to implement corrective measures. We reviewed our experience and identified technical challenges. We propose changes to decrease the number of alerts and alert fatigue without jeopardizing patient safety.

Introduction
Prolonged QT is an independent risk factor for Torsades de Pointes (TdP), a ventricular arrhythmia and a cause of sudden cardiac death. Prolonged QTc is defined as a QTc >470 milliseconds (ms) in men and >480ms in women. For both sexes, QTc >500ms is very abnormal and is associated with a 2- to 3-fold higher risk for TdP. Additionally, an increase of QTc of 60 ms over baseline is considered highly torsadogenic. Prolonged QTc can be congenital (long QT syndrome) or secondary to several medications, electrolytes abnormalities and comorbidities. Because of these risks, national guidelines recommend the use of electronic systems to detect potentially arrhythmogenic QTc prolongation in hospitalized patients. Mayo Clinic has developed and implemented several clinical decision support interventions to identify these high risk patients. In this work, we discuss methods and pitfalls of automated identification of QTc prolongation of >60ms in two consecutive electrocardiograms (ECG).

Methods
A clinical decision support (CDS) rule integrated with the electronic health record (EHR) was implemented to screen all new ECGs to identify an increase in QTc interval >60ms in consecutive ECGs. If QTc prolongation was identified, a semi-urgent alert was sent to the inbox of the ordering provider or primary hospital service. The CDS rule also identified any ECG with QTc >500ms and sent a semi-urgent alert. The generated alerts suggested patient specific modifiable risk factors that may be associated with the QTc prolongation, such as medications and electrolytes. We reviewed all alerts generated by the CDS system related to QTc prolongation > 60ms from 1/1/2010 through 12/14/2015.

Results
The system generated 6,798 alerts in 6,039 unique patients (52.2% male, mean age 60.1 years, SD ± 19.5). Alerts were sent to 6,434 unique providers. The time interval between consecutive ECGs was highly variable: 27.35% of alerts with interval less than 2 days, 23.4% between 2 to 7 days, 34.25% between 8 to 365 days, and 15.0% between 1 year to 5 years. The figure shows the distribution of the alerts based on the QTc intervals of the previous ECG (y) and current ECG (x). Approximately, half of the alerts (46.1%) occurred when both QTc intervals were between 360 and 500ms (B). Few alerts occurred when both QTc intervals were at extreme values (<360ms and >500ms), both low (A=0.1%) or both high (C=2.6%). However, 47.3% of all alerts (C=2.6%+F=44.1%+E=0.6%) occurred with a current QTc >500ms, which would trigger the second alert regardless. Manual review of a sample of ECGs with extreme QTc measurements revealed both accurately and inaccurately measured QTc intervals. Most erroneous measurements observed were attributed to low T-wave amplitude in several leads.

Conclusion
Our preliminary analysis suggests that close to half of the alerts for QTc prolongation >60ms occurred at the same time the current QTc interval was >500ms. Clinically, these patients might be at higher risk of complications and require special attention; however quantification of this potential increased risk requires further study. Two separate alerts may not convey combined risk, if it exists, and may contribute to alert fatigue and alert override. Additional analysis is needed for extreme QTc values found in our samples to identify technical difficulties measuring the QTc interval and implement corrective measures to avoid false positive alerts.
Introduction: In May 2015, Brigham and Women’s Hospital (BWH) transitioned from an electronic health record (EHR) system called the Longitudinal Medical Record (LMR) to an Epic EHR. Both systems provide clinical decision support for drug-drug interactions (DDIs), but each system uses a different knowledge base (KB), user interface, and workflow. The LMR uses a highly-tailored KB to provide only clinically relevant alerts as determined by a panel of experts; the implementation of Epic at BWH uses a commercial KB with minimal customization. In LMR, DDIs are alerted at the time of medication selection; in Epic, DDIs are alerted at the time of order signing after all drugs have been selected. Both LMR and the implementation of Epic at BWH sort drug interaction pairs into three tiers of severity (where Tier 1 is the most severe), although the drugs in each tier differ according to the KB. In LMR, Tier 1 alerts are interruptive hard stops that require the user to cancel one of the interacting drugs, Tier 2 alerts are interruptive alerts that require the user to provide a reason when opting to override an alert, and Tier 3 alerts are passive and provide in-line text that there is a DDI. In the implementation of Epic at BWH, all three tiers of alerts were initially interruptive but could be overridden and providing a reason for the override was optional. After March 1, 2016, in response to concerns about alert fatigue, Tier 3 alerts were filtered so that they are only displayed to users on request.

Methods: Retrospective data was analyzed for the final six-month period in which LMR was in use and for the initial six-month period in which Epic was in use to evaluate alert burden and acceptance rate in each system. Additionally, DDI alert data from Epic was collected for the three months before and after Tier 3 alerts were filtered, in order to measure whether decreased alert burden would increase alert acceptance rates. For each system and time period, all DDI alerts that fired in the outpatient setting were collected, as well as the severity of each alert and the action taken by the user when presented with the alert. Historical medication orders transferred to Epic were excluded, as were orders on test patients.

Results: In LMR, there were 1.1 interruptive DDI alerts per 100 orders. The implementation of Epic had a much higher alert burden, with 10.4 DDI alerts per 100 orders. More concerning, the acceptance rate for drug interactions fell from overall 28.6% to 6.6%. Tier 1 alerts in LMR were accepted 100% of the time (since the system did not allow users to override the alert), but such alerts were accepted only 9.08% of the time in Epic. Tier 2 alert acceptance fell from 27.29% in LMR to 7.81% in Epic. Tier 3 alert acceptance was 5.44% in Epic – there was no comparable figure in LMR, as Tier 3 alerts were informational only and could not be accepted. We also compared the five most frequently alerted interactions for each severity tier in each system and found that acceptance rates fell a large amount for every drug pair compared.

Given the substantial drop in acceptance rates for DDI alerts, the BWH CDS committee hypothesized that the increased alert burden, particularly of Tier 3 alerts, was causing alert fatigue, leading to lower acceptance of Tier 1 and Tier 2 alerts, and decided to filter the Tier 3 alerts. After the change, alert burden fell by 51%, which led to a small increase in acceptance rates. Tier 1 alert acceptance rose from 8.8% to 12.4%, while Tier 2 acceptance rose from 7.0% to 8.8%. Though slight, this difference was statistically significant using Pearson’s chi-squared test for independence in both tiers (P = 0.011 for Tier 1 and P < 0.001 for Tier 2).

Discussion: There are about ten times as many DDI alert firings in Epic as in LMR, and acceptance rates in Epic are much lower than in LMR. There are several possible explanations for the disparity in rates. Each system uses a different KB, and there may be more interactions in the Epic KB that rarely warrant discontinuation of a drug. However, when we limit our focus to specific drug pairs, the difference in acceptance rates persists, suggesting that the difference in KB does not fully explain the difference in acceptance rates. Another possibility is that the increased alert burden in Epic causes alert fatigue. However, our findings show that cutting the alert burden by 51% only caused a small increase in acceptance rates, not nearly enough to make up the decrease in acceptance rates after switching from LMR to Epic, suggesting that alert fatigue may play a role but is not the sole factor. A third possibility is that the greater differentiation between severity tiers in LMR is more effective at convincing users to discontinue drug orders (for good or bad) than the equivalent in Epic.

Conclusion: Analysis of DDI alert data shows higher acceptance rates in LMR than in Epic that are not fully explained by a difference in KB and alert fatigue, suggesting that severity tiering and other workflow differences may be important factors.
Mining EMR Data to Hypothesize Causal Associations for Depressive Disorders

Orhan Abar, M.S. and Ramakanth Kavuluru, Ph.D.

1Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky
2Department of Computer Science, University of Kentucky

Introduction: Mining electronic medical records (EMRs) to identify potentially meaningful and novel clinical associations has been gaining popularity in the medical informatics field. However, it is well known that associations do not necessarily indicate causal relationships owing to confounder variables. If confounders are known, stratification and matching approaches [1] exist that can account for the effect of confounders. In general, identifying confounders is often time consuming and involves significant domain expertise giving rise to automated approaches for identifying and controlling for such variables [2]. We employ such approaches to mine diagnosis and medication data and hypothesize causal associations for depressive disorders using a large EMR dataset of all patient visits (∼ 3.25 million) to the University of Kentucky’s medical center and affiliated clinics during the ten year period 2004-2013. We chose depressive disorders given they form one of the most common mental disorder group, especially among adults.

Methods: To handle sparsity with large numbers of unique codes, we employ grouping techniques. For diagnoses, we use ICD-9 code classes developed by the AHRQ’s Healthcare Cost and Utilization Project (HCUP). These classes group related codes resulting in 282 classes for the 11,877 codes in our dataset. We rolled-up the Multum medication codes using their class hierarchy which resulted in 150 classes (e.g., Penicillins) from a total of 1064 unique medication codes. In each patient visit transaction, we then replaced the codes with the corresponding HCUP and Multum classes resulting in a total of 432 unique variables populating 3.25 million transactions. The depressive disorder HCUP class has a total of sixteen ICD-9 codes representing all variants of depression in ICD-9-CM and occurs in EMRs for 54,923 visits. We first identify and rank potential confounder variables for the HCUP depressive disorder group using L1-regularized logistic regression based feature selection because it has been shown to be comparable to the manual expert propensity score based approach [2]. We repeat this process for all items that have a 95% confidence interval based odds ratio lower bound ORLB > 1 with regards to the depressive disorder group as the outcome variable. That is, we identify potential confounders for all variables that are known to have statistically significant associations with our outcome variable. For each of these variables, we rank-aggregate (using Borda count) to rerank potential confounders in the intersection (given confounders ought to influence both antecedent and outcome variables) of the corresponding sets for the variable and the depression HCUP class. We then build the corresponding so called “fair dataset” [1] that matches patient visit transactions where the top k (for k = 5, 10, 20) confounder variables take on the same set of values and compute ORLB on the fair dataset for the particular antecedent variable. We have also added the well known age-group (based on NLM’s MeSH classification) variable as an additional confounder regardless of whether it was included based on the automated analysis.

Results: We rank-aggregated the antecedent variables that had an ORLB > 1 computed with their corresponding fair datasets by averaging ORLBs from the three scenarios with k = 5, 10, and 20. We had a total of 48 variables remaining after excluding those with support less than the minimum support of 100 co-occurrences with depressive disorders. We hypothesize these variables as having causal associations with depressive disorders (within the scope of the variables, medications and diagnoses, we considered). The top variable (based on the average ORLB rank) is the class of drugs sulfonamides typically used to treat bacterial infections. This is followed by disease group non-Hodgkin’s lymphomas and drug class bronchodilators. There were a total of 313 antecedent variables with support ≥ 100 and ORLB > 1 with depression as the outcome variable when ORLB is computed across the full dataset (that is, before accounting for confounders). This means only 48/313 ≈ 15% of the full set of statistically significant associations can be hypothesized to be causal based on our analysis. Furthermore, out of the top 20 associations from the list of 313, only the drug class of CD52 monoclonal antibodies also occurs as a causal association. Suicide, self-inflicted injury, personality, dissociative, anxiety disorders that have high ORLBs did not end up in the causal variable list after accounting for confounders.

References

Integrating Medical-Dental Care for Patients with Diabetes: A Pilot Implementation of Clinical Decision Support Alerts

Amit Acharya, BDS, MS, PhD1,2, Kelsey M. Schwei, PhD1,2, Dixie Schroeder, MBA1,2, Jordan Ashton, DO2, Srinivas Challa, DMD, MPH2, Cindy Sorenson, FNP2, Louay Danial, MD2, John O’Brien, DDS2, Eric Penniman, DO2

1Institute for Oral and Systemic Health, Marshfield Clinic, Marshfield, WI; 2Marshfield Clinic Health System, WI;

Abstract
While poor oral health impacts overall health, care coordination and integration of medical-dental practices can better address this issue. This study aimed to understand the challenges of implementing clinical decision support alerts (CDS-A) to support care coordination of patients with diabetes at medical and dental centers.

Introduction
It has been established that diabetes has a deleterious effect on gum disease. Evidence is growing regarding the impact of periodontal health on diabetic control. In primary care it is a routine practice to conduct periodic foot and eye exams for patients with diabetes and refer them to the specialist if required. However, conducting regular oral exams for patients with diabetes and referring them to dentists is not standard practice. One out of every four patients with diabetes is unaware of the condition and dental visits can be utilized for screening. The objective of this study was to pilot the implementation of CDS-A within an integrated Electronic Health Record (EHR) to facilitate better care coordination of patients with known/unknown diabetes at the medical and dental centers.

Methods
Two medical and two dental practices at the Marshfield Clinic Health System were targeted for the pilot implementation of the CDS-A. At the medical center, the study team developed a CDS-A which would remind the primary care providers (PCP) of the need to conduct a visual oral examination for their patients with diabetes. The CDS-A would trigger in the EHR for the patients with known diabetes based on the following rules: If a patient was partially dentulous and has not been to a dentist in more than six months; OR If a patient was completely edentulous AND has not been to a dentist in more than 12 months. At the dental center, the study team developed a CDS-A which would inform the dental providers that their patient may be diabetic and a chair-side blood glucose screening is needed. The CDS-A would trigger in the EHR for patients with undiagnosed diabetics based on the following rules: If a patient presents with a Body Mass Index (BMI) > 35; OR If a patient presents with a BMI >30 AND age >60 years. The following data from the pilot sites was evaluated: number of CDS-A triggered (medical and dental sites), number of visual oral examination conducted (medical sites), number of blood glucose tests conducted (dental sites), number of advice provided to see the dentists/PCPs, and number of referrals made to the dentists/PCPs.

Results
During the one year data collection period from the medical sites, there were a total of 2,120 unique patients for whom the CDS-A triggered indicating the need for an oral examination. Of this, 1,455 patients received a visual oral examination by the PCPs, 266 patients were advised to see a dentist, and 77 patients had referrals made to dental providers. The pilot at the dental centers was stopped after a period of 5 months as a result of Clinical Laboratory Improvement Amendments certification issues with the glucometers used. During data collection, there were a total of 160 unique patients for whom the CDS-A triggered indicating the presence of diabetes. Of this, 157 patients received a blood glucose test at the dental center. One patient was referred and two were advised to see a PCP.

Discussion
There was provider and operational reluctance to changes in clinical workflow. The medical center pilot yielded valuable information to continue our efforts toward a system-wide implementation of the CDS-A. The dental center pilot identified several challenges which call for further investigation before system-wide implementation.

Acknowledgement
This study was funded in part by a grant from Delta Dental of Wisconsin and a grant from DentaQuest Foundation.
A National Survey Assessing How Many Records Providers Are Allowed to Open at Once in Electronic Health Records in Hospitals and Ambulatory Sites

Jason Adelman, MD, MS1; Matthew Berger, MD2; Amisha Rai, PA2; William Galanter, MD, PhD3; Gordon Schiff, MD, PhD4 David Vawdrey, PhD1; Robert Green, MD, MPH1, Hojjat Salmasian, MD, PhD1, Ross Koppel, PhD5; William Southern, MD, MS2

1Columbia University Medical Center/NewYork-Presbyterian Hospital; 2Montefiore Medical Center; 3University of Illinois at Chicago; 4Brigham and Women’s Hospital; 5University of Pennsylvania

Background: Placing orders on the wrong patient occurs frequently, as exemplified by one hospital that reported over 5,000 wrong-patient orders in one year (a rate of 58 wrong-patient orders per 100,000 orders).1 Some safety experts argue that Electronic Health Record (EHR) systems that allow multiple patient records to be open at one time increase the risk of wrong-patient errors. However, little is known about the number of records hospitals and ambulatory sites allow their providers to open at once. To address this knowledge gap, we conducted a national survey of Chief Medical Information Officers via the American Medical Informatics Association (AMIA) and the Association of Medical Directors of Information Systems (AMDIS) listservs.

Methods: We conducted an electronic survey of CMIOs to assess the number of records hospitals and ambulatory sites allow clinicians to open at one time. In March 2014, we posted the survey on the AMIA and AMDIS listserves, followed by three reminder postings per listserv spread throughout the month. Respondents were asked to identify their EHR vendor, their facility’s location, the maximum number of records their organization allows providers to open at once, and the rationale for this decision.

Results: 112 surveys were completed, of which 79 organizations had both inpatient and outpatient EHR systems, 25 organizations had only inpatient EHR systems, and 8 organizations had only outpatient EHR systems. Of the 104 inpatient EHR systems, 91 allowed multiple records open; and of the 87 outpatient EHR systems, 76 allowed multiple records open. The inpatient and outpatient EHR systems vendors included: Epic (60), Cerner (48) Allscripts (15), Meditech (12), McKesson (9) Siemens (9) and General Electric (7). The healthcare organizations were from around the country, including 25 organizations from the South, 16 from the West, 32 from the Midwest, and 32 from the Northeast. 44% ±5% of organizations allowed the maximum number of records open, while 38% ±4% restricted providers to one record open at a time (Table 1). Some examples of the rationales provided for how many records organizations allow open at once are given in Table 2.

Table 1. The number of records providers are allowed to open at once by healthcare setting.

<table>
<thead>
<tr>
<th>Healthcare Setting</th>
<th>Max (3 or More Records)</th>
<th>Hedge (2 Records)</th>
<th>Restrict (1 Record)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inpatient</td>
<td>38 (41.8%)</td>
<td>16 (17.6%)</td>
<td>37 (40.7%)</td>
<td>91</td>
</tr>
<tr>
<td>Outpatient</td>
<td>36 (47.4%)</td>
<td>13 (17.1%)</td>
<td>27 (35.5%)</td>
<td>76</td>
</tr>
<tr>
<td>Total</td>
<td>74 (44.3%)</td>
<td>29 (17.4%)</td>
<td>64 (38.3%)</td>
<td>167</td>
</tr>
</tbody>
</table>

Table 2. Survey respondents’ explanations for the number of records providers should be allowed to open at once.

<table>
<thead>
<tr>
<th>Comments from those who allowed three or more charts open.</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The efficiency benefits are such that this is justified. There are other ways to prevent wrong patient errors.”</td>
</tr>
<tr>
<td>“Within our EHR, there are certainly efficiencies associated with having multiple patient charts open at once. It allows the clinician to quickly move between patient charts. We have had very little issues with documentation of orders in the incorrect chart. Thinking back to the paper record, there is really no higher risk than gathering multiple patient charts to complete rounds or documentation.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comments from those who allowed only 1 charts open at a time</th>
</tr>
</thead>
<tbody>
<tr>
<td>“My organization chooses to allow only one EHR open at a time to decrease potential documentation errors by staff documenting on the wrong record, as well as treatment errors by staff reviewing orders and/or results on the wrong patient. We feel, as do the organizations we polled, that multiple records open by the same person is not good practice and is an error waiting to happen.”</td>
</tr>
<tr>
<td>“Totally against having more than one record open at a time for patient safety reasons. I have been at a previous organization that allowed multiple records open and have seen wrong-patient errors occur. It MAY be more efficient - I have no proof of that - but for patient safety the limit should be just a single patient chart open at a time.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comments from those who hedged at two charts open at a time</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Two seems to represent the sweet spot between efficiency and safety as long as training is present to mitigate the risks.”</td>
</tr>
<tr>
<td>“Have to balance efficiency and chance for error (even occurred on paper charts). For ED I think the number should be 4 and 2 for all other venues.”</td>
</tr>
</tbody>
</table>

Discussion: The results of our survey demonstrate there is no consensus on the number of records that should be allowed open at the time of placing orders. In fact, the proportion of organizations that allow the maximum number of records open (44%) is not significantly different than those who restrict providers to accessing only one record at a time (38%) (p=0.32). Interestingly, 17% ±3% of organizations aim to balance the risks of wrong-patient errors and the benefits of increased efficiency by hedging with two records open at once. These results are not surprising, as there is currently no evidence to support a best practice. We are conducting an AHRQ-funded study to assess the relationship between the number of records concurrently and the risk of placing an order on the wrong patient (AHRQ 1R21HS023704).

References:
Introduction

Electronic prescribing promises to reduce legibility and transcription errors during the prescribing process. However, electronic prescribing can also introduce new kinds of errors. When prescribing medications in the Partners Healthcare outpatient Longitudinal Medical Record (LMR), practitioners have the option of writing additional information in one of two fields: “Special Instructions” and “Comments”. Text written in “Comments” usually contains notes to providers and is not viewable by the pharmacy or patient; text written in “Special Instructions” is transmitted to the pharmacy or printed on the prescription. This presents the potential for dangerous errors: providers could mistakenly write important information intended for the patient or pharmacy in “Comments”, thinking that they are communicating with the pharmacy or the patient. We investigated the prevalence of providers inputting information meant for the pharmacy in “Comments” and analyzed what information these comments contained.

Methods

LMR medication record data were requested for Brigham and Women’s Hospital ambulatory care patients from 1/1/2000 to 5/31/2015 (20,123,881 records). From this database, we extracted orders that were electronically sent to the pharmacy where the “Comments” field did not match the “Special Instructions” field or was not auto-populated by information found in the Partners Medication Dictionary (6,060,272 medication orders). From this database of medication orders, a random sample of 10,000 prescriptions were reviewed and manually coded by a research assistant, with results verified by a pharmacist.

Results

10,000 randomly selected comments were reviewed, and 862 (8.6%) of those comments appeared to have been meant for the pharmacy. Examples include: “Pharmacist, please do not fill this rx, it has been changed to lotrimin, thank you” and “please write instructions in Spanish”. We coded these comments using five categories, shown in Table 1. Additionally, we found that 82 (9.5%) of the comments contained information that conflicted with the rest of the prescription and 71 (8.2%) comments contained redundant information.

Table 1. Categorization of comments that were meant to be sent to the pharmacy.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Count (n = 862)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose instructions/clarification</td>
<td>273</td>
<td>31.7</td>
</tr>
<tr>
<td>Other instructions to pharmacist</td>
<td>194</td>
<td>22.5</td>
</tr>
<tr>
<td>Warnings and other instructions to patient</td>
<td>153</td>
<td>17.7</td>
</tr>
<tr>
<td>Substitution information</td>
<td>108</td>
<td>12.5</td>
</tr>
<tr>
<td>Formulation and quantity</td>
<td>95</td>
<td>11.0</td>
</tr>
</tbody>
</table>

Conclusion

While the majority of comments in the LMR “Comments” field are meant for other providers, 862 (8.6%) of analyzed comments appear to have been meant for the pharmacist. This suggests that some providers may not realize that the “Comments” field is not transmitted to the pharmacy and include important information in this field. The system should be reconfigured to ensure that providers know where to put information that is meant to be seen by the pharmacy when sending electronic prescriptions.
The Super Annotator: A Method of Semi-Automated Rare Event Identification for Large Clinical Data Sets

Patrick R. Alba BA\textsuperscript{a,b}, Olga V. Patterson PhD\textsuperscript{a,b}, Benjamin Viernes MPH\textsuperscript{a,b}, Daniel Denhalter BS\textsuperscript{a,b}, Nicole Bailey MPH\textsuperscript{c}, Andrew Wilson PhD\textsuperscript{b,c}, Aaron W.C. Kamau MD, MS, MPH\textsuperscript{c}, Scott L. DuVall PhD\textsuperscript{a,b}

\textsuperscript{a}VA Salt Lake City Health Care System, Salt Lake City, UT; \textsuperscript{b}University of Utah, Salt Lake City, UT, USA; \textsuperscript{c}Anolinx LLC, Salt Lake City, UT

Abstract

Detecting rare events in an electronic medical record (EMR) can be like searching for a needle in a haystack. Automated methods like natural language processing (NLP) and manual methods like chart review can be used, but when events are rare and the patient population is large, manual annotation may not be feasible and NLP alone may not be able to reach acceptable levels of performance. We propose a semi-automated method for NLP-assisted retrieval of relevant documents and manual review of the resulting instances. This method was developed to identify medication-related adverse events among patients receiving care in the U.S. Department of Veterans Affairs.

Background

Success in detection of an adverse drug event (ADE) using EMR depends not only on how often the medication is used relative to how often an ADE occurs, but also on frequency of medication mentions in notes, which defines the search space size, relative to frequency of ADE documentation. For example, given a medication that is 1) administered twice a month, 2) prescribed to 1,000 patients, 3) observed over a one-year period, and 4) associated with an ADE published to occur in 1% of patients, there may be only 10 mentions of ADEs among 24,000 mentions of medication administrations documented in the clinical notes. Creation of a knowledge base sufficient for a fully automated solution requires manual annotation of documented cases. However, the number of clinical documents where ADEs can potentially be mentioned may be too great for manual review of a random set.

Methods

We propose a semi-automated method of identifying drug-related adverse events from the EMR with the following steps: 1) Using relevant drug administration codes from structured data, identify patients who received the target medication during the selected study period. 2) Create a clinical document corpus that contains clinical notes found on the day of administration or with keywords relevant to the medication. These documents potentially have a mention of an ADE. 3) Develop an initial list of keywords or phrases that could identify the ADE of interest and flag documents in the selected corpus that contain one of these patterns. At this point, the majority of mentions may involve drug label warnings, patient instructions, and other instances that do not represent a patient actually having an ADE. 4) Use other relevant words in the flagged examples to expand the list of ADE keywords and phrases. 5) Use the flagged examples to determine a heuristic of how close ADE related keywords have to be to medication-relevant keywords to be associated with the target medication (for some examples, presence of an ADE keyword in the document may be enough). 6) Create candidate snippets of text from the flagged examples and group snippets that contain the exact same phrase. Order the snippet groups by descending frequency and review to eliminate patterns that do not pertain to the patient experiencing an ADE. Reviewing these snippets as groups allows hundreds of instances to be manually reviewed at the same time and potentially excluded as irrelevant from any future iterations of this process. As each snippet group review accomplishes the work of many otherwise individual manual review tasks, it comprises the “super” annotator function. 7) Snippets with lower frequencies or those that cannot be determined as irrelevant without reviewing their individual context can be assigned to the normal manual review process. 8) As ADEs are discovered, any new potentially relevant terms can be added to the ADE keyword and phrase list and steps 4-8 can be repeated until the number of instances is small enough that a detailed manual review of all remaining candidate phrases becomes feasible.

Discussion

Using an NLP-assisted, semi-automated method for detecting rare ADEs can help overcome the combination of having a limited knowledge base to train a fully automated method and having too many documents for manual review. Future work includes applying this approach on a clinical use cases to test its time saving relative to the traditional approach of chart review or natural language processing.

Acknowledgements

This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204. Additional funding support came from Anolinx, LLC through the University of Utah Center for Scalable Analytics and Informatics, a National Science Foundation Industry & University Cooperative Research Program (I/UCRC) # IIP-1439668.
Approaches for using temporal and other filters for next generation phenotype discovery

David Albers, PhD, Adler Perotte, MD MA, and George Hripcsak, MD MS
Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

There are two processes that contribute to health data, human biology and the noisy measurement process, caused by the healthcare process. Additionally, these systems are not static in time; when we account for time, even the measurement process is infinitely dimensional and difficult to manage. Because of these complexities, inferring useful information from electronic health record (EHR) data is difficult, opaque, and yet is necessary. To make the inference more transparent we formulate the relationship between the humans being cared for and the process of observing them explicitly to understand what choices these processes impose about our inferences and to appreciate the consequences our inference choices have on what is observable given an inference method. Specifically, when we attempt to infer a feature about patients given their data, we have four choices, (i) we can transform the data using clinical and biological insights into a targeted form (e.g., a median) and input that into a machine learning algorithm, (ii) we can input raw data directly into a machine learning algorithm, (iii) we can impose clinically and biologically motivated constraints on the machine-learning–optimization-system–identification machine, or (iv) we can try to infer the measurement process, undo its effects, and input the transformed data into a machine learning algorithm. Here we will formulate an explicit framework for the generation of healthcare data and then demonstrate how the projection options (i)-(iv) can impact what we can observe.

Methods

We have two goals, formalizing the EHR-data generating process and demonstrating how inference choices affect what is possible to observe. This framework, motivated by stochastic filtering and inverse problems, explicitly defines the relationship between the data we collect and the underlying true human state we would like to study. Moreover, we use this framework to conceptualize different methods of representing laboratory variables for use in high-throughput computing applications such as phenotyping and cohort identification. Then, to explicitly demonstrate how the different inference pathways (i-iv) can affect what we can infer about the true human state given the data biased by the healthcare measurement process, we apply several analyses including lagged correlations, principle component analysis, Bayesian networks, and Kalman filters to data from the Columbia University Medical Center.

Results

We begin with the framework and follow with inference examples constrained by the framework. Consider for individual i the state, \( x_i \), the biology and environment that map the state forward in time, \( x_i(t+dt) = f_i(x_i,t) \), and the measurement process, \( m \), with noise, \( n \), \( y_i(T+dt) = m(f_i(x_i,t)) + n(m(f_i(x_i,t),t) \), generating all the measurements for an individual, \( \{y_i\} \), or for a population \( Y = \{y_1, y_2, \ldots, y_N\} \). Next introduce a diagnostic function, \( d(y_i) \), such as the median of a patient’s data. Healthcare bias is generated by \( m() \) and \( n() \). Ideally, \( m() \) would be the identity and \( n() \) would be small and independent of \( m() \), but when they are not, what we can derive from data, \( d(y_i) \) may have no relationship to the truth, \( d(x_i) \). In this case, it is very difficult to infer \( f_i \) from raw data. Pathway (i) requires selecting an intermediate diagnostic function \( h(y_i(T)) \) using either clinical knowledge of \( m() \) and \( n() \) or a modeling technique targeted at a reducing a problematic feature of \( m() \) and/or \( n() \) to minimize the bias imposed by \( m() \) and \( n() \) on \( y_i \). Explicitly, we choose \( h() \) such that \( d(h(y_i(T))) \) is very similar to \( d(x_i) \) or captures a desired feature of \( x_i \) effectively reducing the dimensions or bias for inferring a feature of \( f_i \). Examples include removing outliers generated by \( n() \) that can confound \( d() \) or capturing important temporal pairwise relationships between variables. We then use the de-biased output of \( h(y_i) \) in a machine learning algorithm to infer features of \( f_i \). This method could be iterative. Pathway (ii) requires applying machine learning to relatively raw healthcare data, either \( \{y_i\} \) or \( Y \), attempting to directly infer \( f_i \) from the data while ignoring \( m() \) and \( n() \). Pathway (iii) is pathway (ii) with constraints imposed on the data or the model. Pathway (iv) directly attempts to infer \( m() \) and \( n() \) given data and then uses our understanding of these processes to de-bias \( y_i \). To verify the observable effects of choosing one inference pathway over another we compare the resulting analysis of lagged correlations versus Kalman filter, Bayes nets, and PCA. We find many differences. For example, lagged correlations (i) between laboratory variables reveals intervariable relationships, e.g., sodium’s relationship to magnesium that do not appear in the Kalman filter or PCA analyses (ii). Nevertheless, the Kalman filter and PCA reveal interesting intervariable relationships and effectively reduced the dimensions by revealing the most important laboratory measurements that account of the most variance.

Discussion

We show how relationships embedded in data are differently extractable and represented when we make different choices regarding the path to inference. We cannot include every possible data element and encoded temporal relationship because we have neither the data nor the computational power to do so. Because the analysis of a variable affects what we can infer about it, we have make choices about the form of variables included in our analyses. Here we show some of the generalizable effects of different inference pathways, helping to inform goal driven choices when applying data science methods to EHR data.

Acknowledgments

We acknowledge NLM grant R01 LM06910 and NHGRI grant U01HG008680.
Heteroscedastic Omics Data Analysis: Distance-Based Multivariate Approach

Alexander V. Alekseyenko, PhD

1Program for Human Microbiome Research; The Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC

Abstract

Permutational non-Euclidean analysis of variance, PERMANOVA, is a highly effective test for exploratory analysis of multivariate data. This approach however suffers from loss of power and control of type I error when data are heteroscedastic and unbalanced. We provide several solutions to this problem. We demonstrate the desirable type I error and power characteristics of the new tests and develop a R-Shiny package for data analysis with these novel methods.

Introduction

Permutational non-Euclidean analysis of variance, PERMANOVA, is routinely used in exploratory analysis of multivariate datasets to draw conclusions about the significance of patterns visualized through dimension reduction. This method recognizes that pairwise distance matrix between observations is sufficient to compute within and between group sums of squares necessary to form the (pseudo) F statistic. Moreover, not only Euclidean, but arbitrary distances can be used. The PERMANOVA is highly effective in testing the omnibus hypothesis in data with uniform multivariate spread across factor levels. This method, however, suffers from loss of power and type I error inflation in the presence of heteroscedasticity and sample size imbalances. Under some conditions, PERMANOVA is unable to make any rejections or the test rejects all or most of the times without discrimination.

Methods

We develop several solutions to overcome the issues with PERMANOVA in unbalanced heteroscedastic conditions. First we propose, a sub-sampling approach which stabilizes type I error at the cost of reduced power. Next, we derive a pairwise distance-based computation for the Welch t-statistic, which accounts for unequal variances. Finally, we develop a general distance-based multivariate analysis of variance approach statistics for arbitrary k-level factors. We evaluate these methods on simulated data spanning a wide range of possible sample sizes and multivariate spread. In addition, we demonstrate the performance of the methods on sample sizes typical of a discovery type experiment or a small clinical study.

Results and Conclusions

We demonstrate empirically the desirable type I error and power characteristics of the new tests. Specifically, the novel approaches control the type I error at the prescribed rejection level, and do not exhibit anomalous power loss associated with the PERMANOVA. We compare the performance of PERMANOVA and the new methods in reanalysis of two existing microbiome datasets, where the methodology has originated. In these analyses, we demonstrate how the use of PERMANOVA may result in erroneous conclusions, while the new approaches do not. We develop a package using R-Shiny to enable visualization and multivariate analysis of microbiome datasets using PERMANOVA and the newly developed approaches.
Creating electronic clinical decision support tools for management of obese GERD patients in an outpatient setting.
Olga Alexeeva, BS¹, Rena Yadlapati, MD¹, Raj Keswani, MD¹
¹Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Abstract: The aim of this quality improvement initiative was to design and implement a real-time electronic clinical decision support workflow to improve weight loss and symptom score outcomes in obese GERD patients, and to solicit feedback on the effectiveness of this intervention from physicians and patients.

Introduction: The aims of this quality improvement initiative were 1) to design and implement a real-time electronic clinical decision support (CDS) workflow for providers seeing obese GERD patients to encourage referral to a 6-month weight loss counseling program, 2) to assess the impact of the workflow on physician and patient satisfaction and on patient outcomes, and 2) to solicit feedback from physicians on the efficacy of the workflow and intervention.

Methods: We are conducting a prospective pilot study at a single academic medical center targeting obese outpatients with GERD in internal medicine (IM) and gastroenterology (GI) clinics. Over a six-month period (3/15-9/15), an inter-disciplinary team including GI physicians, NM CMIO, health educators, a dietician, IT staff, a nurse practitioner, and a medical student designed an electronic CDS workflow in Epic to identify at-risk patients and automate referral to a targeted weight management program. Patients are identified via a BestPractice Advisory (BPA) utilizing criteria of documented BMI (≥ 30), ICD-9/10 diagnosis code of GERD, and an active PPI prescription. Education for the GI and IM department providers was provided via e-mail in August, and the workflows went live in both clinics in September. Patient enrollment in the weight loss counseling program initiated on 9/15/15 and is ongoing. Patients are also enrolled in a PPI monitoring program with tapered PPI withdrawal and/or PPI optimization as appropriate; data is not yet available on outcomes of the PPI monitoring program.

Results: An interdisciplinary approach to design and implementation of a CDS workflow targeting obese GERD patients for referral to a targeted weight loss program was successful (go-live date: 9/15/15). Unforeseen interference of our BPA workflow with existing workflows in the IM department necessitated adjustments to the BPA and workflow and expansion of our interdisciplinary team to include IM physician stakeholders. Over a 5-month period, 76 obese GERD patients were identified via the BPA and referred to the weight management program; 36 were successfully contacted; 16 enrolled and are participating (44% successful enrollment). Analysis of workflow impact, physician satisfaction, patient satisfaction, and physician feedback is ongoing and expected to be complete by August 2016.

Conclusion: This pilot study demonstrates that an interdisciplinary team-based approach is effective in developing and implementing an electronic CDS tool. Further analysis will reveal the impact of the tool on physician satisfaction, patient satisfaction, and patient outcomes.
Investigating How Health IT Solutions Responded to Information Needs for Fall-Risk Management: A Case Study

Dari Alhuwail, PhD and Güneş Koru, PhD
University of Maryland, Baltimore County, Baltimore, Maryland, USA

Motivation: Managing falls is particularly challenging for home health agencies and affects the outcomes of many patients. As identified in earlier research, home care clinicians are challenged throughout the episode of care by the lack of relevant information critical to fall-risk management (FRM) in the Physiological, Care Delivery, Educational, Social, Environmental, and Administrative domains. Therefore, it becomes important to investigate how health information technology (IT) can better support FRM by addressing the information needs and management issues.

Aims: (a) Understand how the currently adopted IT solutions responded to the FRM information needs. (b) Identify opportunities to customize and modify the currently adopted IT solutions as well as adopting other solutions to better support FRM.

Methods: After obtaining the institutional review board’s approval, a case-study approach was followed in one suburban not-for-profit Medicare-certified home health agency (HHA) with three branches in Maryland. The HHA is a part of a healthcare network including a number of hospitals, physician practices, and rehabilitation centers. The case-study was appropriate because the phenomena of interest were inseparable from their context. A diverse sample of full-time nurses, physical therapists, and occupational therapists participated in semi-structured interviews (n=20) and four focus groups (n=15). The interview and focus group transcripts were analyzed using the Framework method.

Results: To support FRM, the clinicians mainly used (i) the electronic health record (EHR) and (ii) a separate fall reporting application (FRA). (iii) The evidence also highlights opportunities for other IT solutions that are not currently adopted. (i) The EHR was helpful in capturing, processing, and reporting most information in the Physiological, Care Delivery, and Administrative domains such as the patient’s musculoskeletal disorders, brief medical history, and insurance information. However, the EHR sparsely provided information about the patients’ functional status, weight-bearing status, or caregiver contact information. The referral documents received by the HHA and entered into the EHR were mostly faxed, often illegible, and did not have any of the information Social, Environmental, or Educational domains such as caregiver availability, clutter in the home, or patient educational level. (ii) The FRA provided comprehensive information in the Care Delivery, Social, and Environmental domains such as fall-risk assessments, caregiver involvement, and the home organization. Because the FRA was not integrated with the EHR, the clinicians manually entered into the application the relevant Physiological information related to the incident after reviewing numerous screens and notes in the EHR. (iii) The clinicians reported that health information exchange (HIE) solutions can improve the communication of the relevant information in all the domains between the HHA and the referring organization by exchanging Continuity of Care Documents (CCDs). However, it should be noted that presently CCDs have no specific placeholder for patient culture, language barriers, or home environment hazards. Home and body sensors can help gather the patients’ Physiological information about activity levels or perform non-invasive fall-risk assessments and report them on a regular-basis to the EHR for the clinicians to review. The clinicians reported that physician portals can be useful for obtaining Physiological and Care Delivery information. Using physician portals, the clinicians can clarify and obtain further information on FRM-related orders or follow-up on requests for additional services. Telemedicine can assist the clinicians to share and exchange Care Delivery and Educational information with the patients and caregivers such as FRM goals or explaining how a newly prescribed medication affects fall-risks. Telemedicine can also assist in the Social domain by involving the caregivers in the communication of the patient’s progress in care. Patient portals can aid in alerting the clinicians timely to take actions to reduce fall risks when the patients or caregivers self-report negative changes in health conditions or functional status. Patient portals also aid the clinicians in the Educational, Social, Environmental, and Administrative domains by capturing, reporting, and sharing information with the patients and caregivers such as how a health condition affects fall risks, caregivers’ availability, the dangers of clutter in the home, and updates to the patient or caregiver contact information.

Conclusion: To better support FRM, HHAs need to evaluate the current and potential IT solutions’ ability to address the information needs, capacity to store information in a logical manner, and capability to exchange information with other IT systems. Without doing so, health IT solutions may not support FRM.

References


Heuristic Evaluation of a Novel Inpatient Patient Portal
Sana Ali, MPH, BDS, Baria Hafeez, MS, Lisette Roman, BA, Jessica S Ancker, MPH, PhD
Department of Healthcare Policy & Research, Weill Cornell Medicine, NY, NY

Background: The patient portal is a secure online website (1) through which patients can view their medical records and perform tasks such as request medication refills (2). Portals must be well-designed before patients are likely to use them (3). Others have noted that portals may present usability barriers (4), and these barriers might be contributing to the known socioeconomic disparities in adoption rates (5).

Objectives: Contributing to the known socioeconomic disparities in adoption rates (5).

Methodology: We employed a heuristic assessment checklist by Zhang et al (6) to gauge the usability of the patient portal. Three study team members with training in informatics and human-computer interaction examined the portal functionalities, identified usability problems, and assigned a severity score (range: 1-4) to each problem. Agreement was reached in a consensus meeting. Recommendations to improve portal functionalities and suggestions for innovation were conveyed to the development team. The same evaluators conducted a second round of evaluation after selective implementation of the recommendations. The heuristic evaluation was the first phase of a larger project in which patients will be invited to pilot test the site.

Results: The patient portal had an attractive, minimalist appearance and offered an array of potentially useful functionalities, including large quantity of personal medical data and access to family members’ records. In the pre-implementation evaluation, amongst the usability issues identified, the most severe was Failure to use patients' language with a severity score of 4. A striking example of this was that instead of inviting patients to create a report of their medical information, the website invited patients to click a button labeled “CCD export.” Three heuristics, Consistency and Standards, Match between System and World and Prevent Errors each received a severity score of 3. Of the remaining heuristics, 50.0% received a severity rating of 2 and 50.0% a severity rating of 1, indicating minor usability problems and cosmetic problems respectively. The most severe potential error was the availability of an option for patients to accidentally delete the entire medical record within the portal. After implementation of the recommended changes, the heuristic with the highest severity rating i.e., 4 was Failure to use patients’ language. Among the rest of heuristics, 78.5% received a severity score of 1 or 2 and 14.3% received 3. Although the development team followed some of our usability recommendations, they simultaneously implemented new functionality, which introduced novel usability challenges. For example, a new cardiovascular disease module was added to the portal. As a result, the total score remained almost unchanged.

Recommendations: Some of the recommendations made are listed below:

- Reorganize and re-label the navigation menu in a hierarchical way into groupings based on their conceptual underpinnings. A new organization for the navigation scheme was created.
- Rename report creation options with self-explanatory labels and organize the options for report in a thematic/alphabetical order.
- Notify users that any information that they add to this record is not automatically pushed to the hospital EMR and therefore would not be viewed by the physician.

In addition to these, suggestions were also made for improvement and innovation. For example, organizing information by encounter, displaying medications, diagnoses, and other information in a timeline format, creating progressive levels of complexity to make onboarding easier for first-time users.

Conclusion: Heuristic evaluation yielded valuable information to inform the design and development team. The findings underscore the fact that usability evaluation must be done iteratively and concurrently with the rollout of new functionalities. Otherwise the overall patient experience will continue to be impacted by usability barriers.

References:
A Toolkit that Standardizes Data Retrieval across Multiple Perioperative Data Sources

Frank Aline, BS1, Maxim A. Terekhov, MS, MBA1, Teus H. Kappen, MD, PhD1, Karen Y. McCarthy, MS1, Hongjuan Z. Blazer, PhD1, Stephen P. Baker, AS1, Jesse Ehrenfeld, MD, MPH1-4, Jonathan P. Wanderer, MD, M.Phil.1,2

Departments of Anesthesiology1, Biomedical Informatics2, Surgery3, and Health Policy4
Vanderbilt University Medical Center, Nashville, Tennessee, USA1

Introduction

The increasing utilization of Electronic Medical Record (EMR) yields an abundance of clinical data that can be leveraged in a healthcare setting. Combining patient data within the EMRs can answer various clinical research and quality assurance questions1. While some hypotheses can be tested with data retrieved through EMR reporting systems, data to address more complex questions are typically retrieved through unstandardized, custom-built queries. However, research projects2 may share a similar ‘data pipeline’, which can result in unnecessary analytic work, which is time-consuming and costly. Moreover, each custom query varies per developer, which may compromise reproducibility.

The toolkit framework

We standardized information retrieval through a data pipeline implemented in a toolkit using Microsoft Visual Studio Ultimate 2012, Visual C# 2012 and SQL Server Management Studio (Microsoft, Redmond, WA). This application enables the process of translating a question into a query to retrieve the research data in an automated way. The toolkit allows external data to be imported with a high degree of flexibility into the pipeline. The toolkit then uses the external information to select an appropriate patient cohort (Figure 1). The logic determines whether or not procedural or hospital-encounter information is requested. Depending on which common identifier(s) are presented in the initial cohort, the toolkit will then find and validate other identifiers that link patient demographics to hospital-encounter and procedure information. The next step is a selection process where the user can choose additional variables for retrieval (e.g. comorbidities, drugs, fluids, etc.) and aggregation if applicable (minimum, maximum, average, etc.). Finally, a summary report and a log of applied standardized queries are generated.

Comparing the toolkit to the previous approach, the differences were subjectively noticeable, but were not objectified.

Figure1. Workflow visualization of the Toolkit.

Conclusion:

Implementing a standardized interoperable toolkit, that leverages various sources, permits for the use of a uniform approach. This may reduce tedious and costly efforts towards repetitive data requests.

References

Identifying Discrepancies in Diabetes and Diabetic Foot Ulcer ICD-9 Codes Via Administrative Data in a Veteran Population

Latricia L. Allen, DPM, MPH, Eni Njoh, MPH, Gail Powell-Cope, PHD, Tatjana Bulat, MD, Dezon K. Finch, PHD
HSR&D Center of Innovation on Disability and Rehabilitation Research, James A. Haley Veterans Hospital, Tampa, FL

Introduction
Diabetes and associated comorbidities have been ranked the 7th leading cause of death. Prevention programs have been noted to reduce the likelihood of limb loss but unfortunately are not widespread. Thus, amputations and other adverse events related to diabetes is still prevalent. Diabetic foot ulcer may pose a safety risk for adverse events. The number of undiagnosed cases of diabetes is so prevalent, thus we conducted an evaluation to determine the prevalence of adverse events related diabetes and diabetic foot ulcer in a local Veteran population.

Methods
A sample of all patients with diabetes and diabetic foot ulcer ICD-9 codes during the years of 2009-2014 were examined and from those patients with diabetic foot ulcers we checked to see if diabetes was coded. 41,324 patients were noted to diabetes diagnosis code, 3200 patients were noted to have a code for diabetic foot ulcer with diabetes, and 2783 patients with diabetic foot ulcer did not have a diagnosis code for diabetes. Adverse event data are still being currently collected for this project.

Results
41,324 patients were noted to have a diagnosis code for diabetes, 3200 patients were noted to have a code for diabetic foot ulcer with diabetes, and surprisingly 2783 patients with diabetic foot ulcer did not have a diagnosis code for diabetes. The number of patients with diabetic foot ulcer without a diabetes diagnosis code brings questions to mind such as possible false positive or false negatives for diabetes as well as diabetic foot ulcer coding in this patient population.

Conclusion
The data set from this project will be further examined in a chart extraction study to determine any variation in documentation of diabetes and diabetic foot ulcer in the electronic medical record and ICD-9 codes. This information may prove valuable in possibly improving EMR accuracy as well as improve management of patients with diabetes.

References
Towards Automated Selection of Patient-Specific Education Materials in Ambulatory Care Settings

Fatemah M. Aloudah, MS, Janusz Wojtusiak, PhD, George Mason University, Fairfax, VA

Introduction

In the era of personalized health and patient-centered care, providing personalized education materials to patients becomes vital especially for patients with chronic diseases where self-care management is highly required for successful health outcomes. However, during an office visit, fully educating and informing patients is not feasible due to several factors, including short visit time, patients’ varying health literacy levels, and complexity of the information needs to be delivered to patients. Information systems, including certified EHR systems, may be able to remedy this situation by aiding providers in the selection of the education materials to be distributed to patients during or after discharge. One of the effective ways to engage patients in their health, trigger a behavior change or improve compliance to a treatment regimen is by providing personalized education materials. We propose to personalize education materials along five patient-specific dimensions: demographics, diagnoses, prescriptions, health literacy level, and preference of material delivery method/media.

Methods

We reviewed available literature and investigated leading EHR systems to explore what options are available for selection of patient education materials. In the description below we purposefully left out names of vendors. Literature Review: A number of papers have been published since 1990s in regards to generating tailored education materials to patients. Mainly, these earlier initiatives were not implemented in clinical settings, but some were tested in controlled clinical trials. The literature varied between reporting on new systems, case studies and system evaluative studies. Current EHR approaches: We have reviewed options available in five leading EHR systems as part of demonstrations given by the vendors. While the systems differ in how the education materials are selected, there is no real individualization of selected materials. All of the systems allow for selecting materials based on diagnoses for a given patient, and don’t seem to consider wider patient context. The systems either search available resources by keywords defined by ICD-9/ICD-10 diagnoses, or link directly to pre-defined resources for these diagnoses. Sources of education materials: EHR systems link to publically available materials, such as Medline Plus, or to sites of commercial content providers. Among the reviewed EHR systems, one included selection of basic built-in materials, and additional ability to link to more advanced ones by an external provider. Some of the content providers offer materials in multiple languages (mainly English and Spanish, but also others), with varying availability of specific topics. Additionally, some content providers specialize in multimedia content, or in mobile delivery format.

Proposed Solution

The selection of education materials can be done in an automated fashion, with only final approval by clinicians. Artificial Intelligence can be used to match patient characteristics with specific content to be provided, and embed the materials into patient’s context, not simply provide diagnosis-specific information. Such a system should allow for full individual customization of materials based on the 5 dimensions: demographics, diagnoses, prescriptions, health literacy level, and preference of material delivery method/media. The solution requires creation of education materials in small modules that can be assembled in multiple ways to match patient’s characteristics and preferences.

Conclusion

The main identified issue with the current approach is that education materials are not patient-specific, but rather diagnosis-specific. The use of patient information should allow for selection of truly individualized education materials, and narrowing the gap between currently used approaches and what is possible.
Complexity Reduction and Visualization of RDF Knowledge Networks for Precision medicine

Zainab Al-Taie, MS¹, Nattapon Thanintorn, MS², Ilker Ersoy, PhD², Richard Hammer, MD², Dong Xu, PhD ¹,³, Trupti Joshi, PhD ¹,³,⁴, Dmitriy Shin, PhD ¹,²,³

¹Informatics Institute, ²Department of Pathology and Anatomical Sciences, ³Department of Computer Science, ⁴Department of Molecular Microbiology and Immunology, University of Missouri, Columbia, MO, 65211, USA

Introduction
Precision medicine is one of the most rapidly emerging areas of research and development, proving to be crucial for improving patient care, but there is a lack of a comprehensive set of tools that are easy to use for analysis and incorporating genomic data into clinical decisions. In our study, we attempt to reconstruct interrelationships among biomarker proteins, diseases, and signal transduction pathways for individualized treatments. Towards this, we have developed a suite of tools, which can shed some light on tumorigenesis and genomic changes taking place in individual patient.

Method
We have developed a suite of tools for visualization and curation of KEGG pathway data to incorporate it into precision medicine. The curated pathways have been converted into Resource Description Framework (RDF) to create a knowledge base network¹. Our Complexity Reduction and Visualization (CRV) tool reduces the complexity of the knowledge network by finding the more relevant paths among a set of start and end genes. This tool creates three different networks from cohort patients; the first network is created from non-metastatic cancer data, the second network is created from regional metastasis cancer data, and the third one is created from distance metastasis cancer data. The start and end genes list for each network differs; for the non-metastasis network, they are the surface receptors as start genes, and differentially expressed genes in cancer patient as end genes. In the regional metastasis network, they are differentially expressed genes in cancer patients with non-metastatic cancer as start gene set, and differentially expressed genes in regional metastatic cancer after excluding the genes that exist in the start gene set as end gene set. In the distance metastasis network, they are the differentially expressed genes in patients with regional metastatic cancer after excluding the genes that exist in non-metastatic cancer as start genes, and differentially expressed genes in distance metastatic cancer after excluding the genes that exist in the start gene set as end genes set. For individual patient, a network can be created from these three networks in order to get most relevant network to a particular disease. Patient’s disease network after mapping to the three networks can be used by pathologists to predict if a particular patient may develop metastasis. The resulting networks have been visualized using d3.js.

Results
The tool will find the shortest paths among the set of start and end genes to create less complex network than the input network. We have applied this to Lung Adenocarcinoma (LUAD) samples. From the patient cohort dataset, we have found 90 genes and 15 pathways such as focal adhesion, mTOR, and ErbR signaling pathway that have a role in transforming the cancer from non-metastatic cancer to be a metastatic cancer. This set of tools can be applied to answering diverse questions including getting a better understanding of genomics mechanisms that play a role in metastasis vs. non-metastatic cancers and finding the genes that can cause the cancer to metastasize in an individual patient.

Conclusion
Our study represents a novel computational method in pursuing the understanding of interplay between pathways in metastatic cancer and how primary tumor will develop metastasis. The CRV Tool can help in finding key genes, and hypothesis generation. The two major components of CRV which are the complexity reduction and visualization provide pathologists a simplified network of relevant genes and pathways for the individual patient.

References
Designing an evaluation methods course in health informatics
Jessica S Ancker, MPH, PhD, Stephen Johnson, PhD
Weill Cornell Medical College/Weill Cornell Graduate School of Medical Sciences, NY, NY

Background: Health informatics is a multidisciplinary field, integrating aspects of engineering, computer science, medicine, statistics, epidemiology, social sciences, and cognitive science. As a result, health informatics graduate programs may feel fragmented for students, who may encounter multiple perspectives with little sense of a unifying framework. Educational research demonstrates that learning, particularly among adults, is facilitated when students clearly understand instructional goals, relevance, and relationships within the curriculum. Although much excellent discussion has focused on what should be taught in an evaluation methods course, it has not to date focused on the best way to sequence instruction to avoid the perception that the course is a laundry list of methods.

Objective: To propose a unifying framework for an evaluation methods course, which can help students understand the rationale for, relationships between, and appropriate ways to apply a variety of evaluation methods.

Course design: Hierarchical sequencing is a highly effective curriculum design approach; in this approach, knowledge and skills introduced early in training are built upon as more complex competencies are introduced later. However, it is impossible to develop a simple hierarchical sequence for a multidisciplinary curriculum that incorporates methods from several disciplines. Instead, we propose a content relationship structure after a content analysis of evaluation methods. We based the structure upon the Delone and McLean model of information systems success. The structure allows us to group and sequence theories and methods according to underlying constructs.

Course content and competencies: Because the DeLone and McLean model recognizes that information quality, system quality, and service quality are important precursors to information system success, we begin our course with an information quality module focusing on evaluating data quality and related constructs. The module introduces measurement theory and evaluation competencies including test characteristics, information retrieval metrics, and expert rater approaches.

The DeLone and McLean model progresses to user satisfaction, intent to use, and system use. The evaluation course therefore progresses to a users module, which introduces behavioral theories of use such as diffusion of innovation and task-technology fit, as well as evaluation competencies such as survey design, survey data analysis, and psychometric methods.

The next module on human-computer interaction focuses on DeLone and McLean’s system quality, use, and user satisfaction constructs. This module exposes students to cognitive and information processing theories along with methods such as cognitive walkthroughs and usability analysis. A sociotechnical module orients students to the subjectivist perspective and develop competencies in qualitative methods including conducting interviews and focus groups and analyzing qualitative data.

The course concludes with an outcomes module focusing on the DeLone and McLean net benefits construct. This module introduces theories and methods appropriate for assessing benefits as well as unintended consequences, in terms of healthcare quality, patient safety, and cost. One important focus is on approaches drawn from health services research that are useful for informatics interventions, including randomized trials, retrospective cohort studies, and interrupted time series analyses. The other focus is on approaches drawn from engineering such as such as iterative design and evaluation cycles.

Conclusions: Evaluation methods are critical components of graduate informatics training, but the heterogeneity of health informatics makes it challenging to select and sequence competencies. Our curriculum provides context for evaluation methods via a well-validated model of information system success. This curriculum helps students understand the relevance of the methods, how to select methods to address particular questions, and how all methods support the long-term goal of developing effective informatics tools. Furthermore, by introducing methods in the context of this model, we reinforce the applicability of theoretical models to guide research, design, and implementation in health informatics. In our program, we offer the evaluation methods course in the first semester of the master’s program, concurrent with introductory courses on health informatics, biostatistics, and the healthcare system.

References
3. Ammenwerth E, Craven CK, Georgiou A, Mantas J. Health IT evaluation in health informatics curricula: International overview and recommendations. Paper presented at: Medical Informatics in Europe (MIE) 2014;11-01 00:00:00 2014; Istanbul, Turkey.
Configurable Permissioning to Address Diverse Data Collection Needs

Beth Anderson, PhD¹, Charles Tirrell¹, Henry Agnew¹, Frank Farach, PhD¹,
Leon Rozenblit, JD, PhD¹

¹Prometheus Research, New Haven, CT, USA

Abstract

Multi-study research clinical data repositories (CDRs) require complex permissioning schemes to meet IRB and HIPAA requirements. CDRs that push permissioning configuration into menu options increase administrative overhead and the risk of process and quality errors. We describe how analyst-configurable workflows and permissioning structures, instead of one-size-fits-all approaches, can simplify tasks and improve data quality.

Introduction

Organizations that manage multiple studies within the same clinical data repository (CDR) must meet the permissioning and regulatory requirements of HIPAA and their local IRB. Rules for data sharing often differ even between departments within the same institution, requiring complex permission-management capabilities. Lab-specific workflows such as using a central recruiter versus direct study enrollment are common. Research CDRs such as REDCap² handle this variability by providing over a dozen study level permission options.

Methods

In 2015, Prometheus Research restructured their open source research CDR, RexStudy, to take advantage of newer technologies. As part of this restructuring, RexStudy 4.1¹ permissions were initially broken into 10 different options (e.g., enroll participants, enter data, reconcile data, explore data) where users could be granted permission to zero or more of the options. Several RexStudy clients found their workflows were hampered by being broken into too many discrete permissions and requested customizations to have our technology match their internal workflows, each of which differed slightly. Prior research suggests matching computer and human workflows enhanced acceptance of the technology.³ To better understand the most common client needs, we analyzed permission combinations for 243 users of RexStudy across 14 studies in 6 different labs and 3 different sites.

Results

Of the over 3 million possible unique permission combinations available in RexStudy 4.11, clients used only 20. The four most common were: 1) lab admin only (18.5% of users) which allows adding other users to the lab, creating studies, and assigning study permissions; 2) day-to-day study operation permissions (14.8%) which included recruiting, enrolling, entering data, reconciling and exploring data; 3) lab admin plus day-to-day study operation permissions (10.7%); and finally 4) no permissions (11.5%). Based on these results, it was clear users do similar tasks via different workflows and permission restrictions. To improve the permissions and workflows in RexStudy 4.12: (1) analysts can now define roles and (2) assign a workflow to one or more roles. In the one-size fits all approach of RexStudy 4.11, the workflows were disjointed with each task and permission utilizing a different workflow. Using RexStudy 4.12, a semi-technical business analyst can configure a workflow to include all necessary tasks enhancing data integrity by minimizing the chances of missing a step and improving researcher productivity.

Conclusion

Breaking up workflows across permission boundaries can result in a disjointed user experience in a research CDR and may increase the chance of data collection and/or entry errors. By defining the locally-recognized roles needed for a specific research environment, and by connecting role permissions directly to workflows, our new approach minimizes the chance of errors, and improves the overall user experience.

References

1. https://demo.rexdb.org
2. http://www.project-redcap.org
Log File Analysis of User Engagement with Maternity Information Access Point

Adriana Arcia, PhD, RN\textsuperscript{1}, Eva Rose Asaan Warner, BS, BA, RN\textsuperscript{1}
\textsuperscript{1}School of Nursing, Columbia University, New York, NY

Abstract

Maternity Information Access Point (MIAP) is a mobile-ready, web-based patient education platform populated with curated, evidenced-based, multi-media content for maternity care. It is the research version of Care Guide™ by Maternity Neighborhood™. The purpose of MIAP was to evaluate the feasibility and acceptability of the platform for Medicaid-enrolled women. In this study we sought to gain insight from users’ browsing behaviors by analyzing log files.

Introduction

Maternity Information Access Point (MIAP) is a mobile-ready, web-based patient education platform populated with curated, evidenced-based, multi-media content for maternity care. It is the research version of Care Guide™ by Maternity Neighborhood™. The purpose of MIAP was to evaluate the feasibility and acceptability of the platform for Medicaid-enrolled women. In this study we sought to gain insight from users’ browsing behaviors by analyzing log files.

Methods

Platform users were 16 English- and Spanish-speaking Medicaid enrolled women in the second or third trimester of pregnancy. Users were provided with a mobile hotspot with unlimited data to ensure Internet connectivity. Users were invited to log in to MIAP as frequently as desired over the course of four weeks with a request that they do so at least once a week. Users were also invited to message the researchers (as a proxy for their care provider) with any pregnancy health questions they had. Links to site resources were sent via email once a week to promote engagement.

We examined log files of the activity of the users. Interactions with the site were sorted by user and grouped into sessions (clusters of activity). We analyzed: a) number of sessions, b) number of messages sent, c) total number of resources accessed, d) average number of resources per session. Planned analyses include the relationship of the user’s gestational age to the types of resources accessed.

Results

Users logged between zero and seventeen sessions over the course of their participation. Most sent no messages; the remainder sent only one or two. The total number of resources accessed ranged from 0 to 55. Of the users who accessed resources, their average number of resources per session ranged from 1 to 3.8.

Discussion

Users varied substantially in their engagement with the platform. Avid users logged in repeatedly and accessed numerous resources. However, messaging with the nurses running the study was a poor proxy for messaging with the care provider and there was very little uptake of this option.

Conclusion

Analysis of log files can yield valuable insights about users’ engagement with a site. Potential future research includes comparison of popular versus unpopular resources in an attempt to identify structural factors that affect browsing behavior.

Acknowledgments

Columbia University Provost’s Grants Program for Junior Faculty Who Contribute to the Diversity Goals of the University
Nursing Documentation Redesign: A Collaborative Approach

Deborah Ariosto, PhD, MSN, RN
Vanderbilt University Medical Center, Nashville, TN

**Background:** In response to staffing ratio changes, a team of clinicians and industrial engineers were deployed to observe and interview nursing staff to identify opportunities to save nursing staff time. Top priorities identified by staff nurses were unnecessary/redundant documentation, finding equipment, and better communication with the care team. In 2014, executive leadership endorsed the formation of cross enterprise clinical workgroups to address these priorities. This presentation focuses on the methodology applied to a highly successful documentation reduction initiative.

**Methodology:** Over fifty clinical staff from acute and critical care in both the adult and pediatric hospitals was challenged to standardize and simplify nursing documentation. Exclusions included Obstetrics, Neonates, and Psychiatry. A representative core group of nurses met weekly to shape a common documentation model using a terminology framework that reflected professional nursing practice endorsed by the American Nurses Association\(^1\).

The Clinical Care Classification System\(^2\) (CCC) was used to organize nursing care concepts into 15 Care Categories to create a high level framework for nursing centric problems (e.g., acute pain), interventions (acute pain control), and goal setting. SNOMED CT\(^3\) concept matching was used for signs and symptoms and intervention detail. Free text narrative statements captured the overall patient response to care and communication/event notes.

Other strategies included a modified charting by exception approach and elimination of most copy/paste functions. Assessments auto-populated the plan of care and easily visualized the daily progress toward problem resolution. Documentation that did not reflect professional practice or support high value decision support was discouraged.

**Outcomes:** Several pre/post evaluation approaches were utilized. These included timed classroom simulations and real-time clocked documentation times on pilot units. Timed studies showed on average, > 25 fewer minutes spent documenting and a 30% decrease in charted elements overall.

Staff has had an overwhelmingly positive response in that they have more time to spend with their patients and take needed breaks. Rounding during implementation reassured leaders of the value of their investment in staff time and resources and gave staff an opportunity to show their work and tell their stories. One poignant story shared was: “I now had time to give my patient a foot rub when morphine didn’t work”.

**Conclusions:** Providing nurses with a nursing terminology framework that they could shape and own was key. While the informaticists knows the value of standards for data re-use, interoperability, decision support, and system stability, the clinician knows their own practice and unit needs. Working together, documentation redesign achieved both goals.

---

\(^1\) [http://www.nursingworld.org/EspeciallyforYou/StudentNurses/Thenursingprocess.aspx](http://www.nursingworld.org/EspeciallyforYou/StudentNurses/Thenursingprocess.aspx)

\(^2\) [http://www.sabacare.com](http://www.sabacare.com)

Technological Barriers to Situational Awareness in Laboratory Testing

Argus Athanas, BS; Molly Kantor, MD; Meghan Sebasky, MD; Robert. El-Kareh, MD, MS, MPH
University California San Diego; San Diego, CA

Introduction
Laboratory testing involves many team members and is an integral part of clinical care in the hospital. Ineffective information flow can lead to poor situational awareness amongst team members regarding the status of previously ordered tests as well as unnecessary repeat testing and loss of samples. Motivated by specific concerns raised by clinicians in our hospital, we sought to characterize the entire laboratory testing process with the goal of highlighting areas of breakdown and providing context to improve the effectiveness and efficiency of clinical care.

Methods
We interviewed team members from one hospital at UC San Diego Health System, including physicians, nurses, phlebotomists, a lab director and IT staff. We mapped the laboratory testing process from end-to-end (Figure 1), which allowed us to identify several barriers to effective information flow.

![Figure 1. Clinical laboratory testing process and information flow diagram](image)

Results
We identified several steps of the laboratory testing process that created barriers to situational awareness. These barriers included: limited ability to update test status in the EHR, obstacles to correct erroneous collection modes, multiple methods of collection, nursing work-arounds that increase risk of mislabeled specimens, and limited support for handoffs from one care setting to the next.

Conclusions
A thorough mapping of our entire laboratory testing process provided important insights into system-level improvements that can be implemented to improve situational awareness amongst team members and improve the safety and reliability of our testing process.
Standardization of FDA Adverse Event Reporting System to the OHDSI Common Data Model

Amelia J. Averitt, MPH1, Adler Perotte, MD, MA1
1Columbia University, Department of Biomedical Informatics, New York, NY, USA

Abstract

Adverse drug events (ADEs) are responsible for hundreds of thousands of injuries and deaths every year. The FDA Adverse Event Reporting System (FAERS) is a database for the collection of ADEs, but the data is unstandardized and inaccessible to most. To facilitate research with this data, we developed a methodology to extract, transform, and load the FAERS data into a standardized data model developed by the Observational Health Data Science and Informatics (OHDSI) collaborative.

Introduction

Post-marketing adverse drug events have shown that pharmaceutical agents can be harmful or fatal to consumers, despite rigorous clinical safety testing.1 The Food and Drug Administration (FDA) created an adverse event reporting system (FAERS), to capture possible post-marketing AEs and the potential causal drugs. Though the FAERS data is available to the public and accessible through the FDA website, the data requires substantial preprocessing and is unstandardized, making mapping to medical terminologies arduous and research results difficult to reproduce. The Observational Health Data Sciences and Informatics (OHDSI) is an international, interdisciplinary collaboration to support large-scale data analytics. Data standardization is a principal concern of OHDSI, as a common data model and standardized vocabularies necessarily underlie successful collaborative research. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) provides standardized tables for observational health domains including condition occurrences and drug exposures among others. The CDM has previously been utilized to transform and house administrative claims and EHR data, but has not yet been applied to spontaneous reporting data, such as the FAERS data. The transformation of the FAERS data into the OMOP CDM would facilitate research of drug-adverse event relationships and may further support collaborative research that would employ this data.

Methods & Results

The data was standardized according to the OMOP Common Data Model using an extract, transform, and load (ETL) procedure. FAERS data from 2013-2015 was downloaded from the FDA Adverse Event Reporting System website. This data is organized into several files, and consists of seven domains - Demographic, Drug, Reaction, Outcome, Reporting Sources, Therapy, Indication - reported quarterly for every year. The primary unit of observation for each of these tables is the caseid, which represents a single individual, with potentially more than one adverse event, and multiple drugs that may be responsible. After the initial extraction, the data was processed to combine like-tables across all years and quarters. To populate the OMOP CDM, the FAERS data associated with each domain required a series of transformations. The FAERS Demographic table required the least data transformation where birth_year was calculated from the FAERS' reported age at event and event date to populate the OMOP Person table. For the purpose of this ETL, we assume that each caseid represents an individual with a single visit. Therefore, the Demographic table was also mapped to the OMOP Visit_Occurrence table, which captures the time frame of the potential adverse event. The raw data from the FAERS Reaction and Indication tables are recorded as MedDRA concepts, and both tables were mapped to the OMOP Condition_Occurrence table, with the drug indications recorded as primary conditions (58.7% distinct conditions mapped), and the adverse events recorded as secondary conditions (50.8% distinct AEs mapped). The principal challenge of this ETL was the successful mapping of the raw FAERS drug data to systematized concepts. This process was complicated by the free-text entry of the users’ drug use, which was subject to inconsistencies in preferred drug name, dosage, and misspellings. In this exercise, FAERS Drug data was mapped to the OMOP Drug_Exposure table using the NLP tool, MedXN, to process the free-text drug names to RxNorm codes (69% match rate). NLP tools are a frequently utilized when mapping FAERS drug data, but to our knowledge, this is the first utilization of MedXN for this task.

Conclusion

The FAERS data is an important resource for data-mining for adverse events. An ETL of this data into the Common Data Model is a useful standardization of the data, which will support analytics into drug-adverse event relationships. Though this process only addressed the most recent FAERS data, future work will extend this process to all available FAERS data and improve the completeness of concept mapping.

References


1334
Lessons learned from adopting RTLS(Real-Time Location System) based asset tracking system in a tertiary hospital

Hyunyoung Baek, Seok Kim, MPH, Eunhey Kim, Eunja Jung, Rongmin Baek, Hee Hwang, MD, PHD, Sooyoung Yoo, PHD
Center for Medical Informatics, Seoul National University Bundang Hospital, Korea

Abstract

This study aims to share insights and considerations of adopting Real-Time Locating System (RTLS)-based asset tracking system in a medical environment, based on our experience of implementing and operating it in a tertiary hospital setting.

Introduction

The health care industry has seen robust adoption of RTLS technology, generally for tracking medical staff, patients and assets to improve patient safety and efficiency.

Methods and Results

We developed an asset tracking system and an application that is conducive to tertiary hospitals. Bluetooth sensing beacons and tracking tags utilized in the system were also tailor-made. The entire project, from product design to implementation, lasted 13 months starting from November 2013. Five types of devices (i.e. iPad, Patient Monitor, Bladder Scan, Infusion Pump, Oxygen Holder) expected to benefit the most from the system due their high mobility were identified. 4 iPads, 160 Patient Monitors, 15 Bladder Scans, 200 Infusion Pumps, 21 Oxygen Holders were attached with active tags for tracking location. The existing Wi-Fi infrastructure was used for recognizing the asset’s location in the general wards, whereas, the new Bluetooth beacon with higher accuracy in locating asset position were used in the Emergency Center.

A questionnaire was designed to survey user perception of the RTLS-based asset tracking system 3 months after initial implementation. The user satisfaction scale consisted of five factors: information quality, system quality, device type, expected outcome and overall satisfaction. The survey was disseminated to the nurses of 3 general wards, 3 ICUs and the ER and 117 responses were collected. The score of the overall satisfaction ranged from 2.7-3.4 out of 5, indicating that most nurses were satisfied with the service. Most nurses responded positively scoring 3 or higher out of 5 for the helpfulness and their desire for continued use of the service. There were suggestions regarding the adjustment of tag size and the need for facilitating user training programs.

The following are issues and considerations in terms of development and usability of the said system.

1) Scope of asset tracking: Adequate discussion should precede development to accurately define end-users and target assets to drive needs. 2) Locating & positioning: Proper sensing techniques matching the target asset and area should be selected since Wi-Fi and Bluetooth techniques are generally known to have margin errors of 10m and 5m, respectively. 3) Design of Active Tag: The needs of onsite staff and policy guidelines of the organization should be appropriately balanced when designing and adopting a tag frame. 4) System functionality & Usability: User-friendly interface meeting the needs of end-users should be developed. 5) Education & Operation Support: Education is necessary for end-users to use the system smoothly. Supporting or maintenance staff for the continuous operation of the system should be considered (e.g. changing batteries of beacons and tags periodically).

Conclusion

The successful adoption of a RTLS-based asset tracking system in a hospital setting requires accurate identification of targets (both user and asset) based on their work environment analysis. Moreover, technical aspects such as availability of smaller RFID tags and enhancement of sensor accuracy is equally important and necessary.

Acknowledgement: This work was partly supported by the IT R&D program of SNUBH and SKT
End-user Customization of Electronic Medical Record System Provides Measurable Improvements in Documentation Compliance

Anindita Bagchi, M.Sc., RHIA, Jayne Mitchell, ANP-BC, Jiri Sklenar, Ph.D., James O. Mudd, M.D.
Knight Cardiovascular Institute, Oregon Health and Science University, Portland, Oregon

Abstract
A commercially available EMR was optimized for documentation and delivery of evidence-based care for heart failure patients using American College of Cardiology (ACC) heart failure performance measures. The new workflow resulted in an increased frequency of clinical documentation for ACC heart failure performance measures.

Introduction
Current EMRs provide technical means for optimization to specific local needs. Such modifications, to be successful, require a coordinated effort of clinicians and IT specialists.

Hypothesis
Practicing evidence-based medicine in heart failure patients requires clinicians to consider, and subsequently record, a well defined set of variables, e.g. the seriousness of the disease, or whether recommended therapies were provided to the patient. We hypothesized that consolidation of relevant data elements from several locations into one intuitive, interactive form will result in increased documentation compliance.

Results
We observed consistent increase in reporting frequency for each performance measure (Figure 1) after the form was activated

Conclusion
Commercially available EMRs can be incrementally improved in response to specific local needs. Using this tool we improved clinical documentation practices. We enabled this new form initially only for heart failure providers, and developed a road map to include all cardiology providers in Phase 2, and internal medicine and family medicine providers in Phase 3. We expect the new tool will result in better clinical documentation, enhanced use of evidence based heart failure therapies, and reduce variation in heart failure care throughout the organization.

Figure 1. An example of data visualization of frequency with which “ACE/ARB Prescribed” therapy was reported over time demonstrates the impact of system changes.
On the Predictive Potential of Graph Patterns for Biomedical Relation Extraction

Gokhan Bakal, M.S. 2, Sergei Wallace2, and Ramakanth Kavuluru, Ph.D.1, 2

1Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky
2Department of Computer Science, University of Kentucky

Introduction: Biomedical relation extraction (BRE) is a crucial information extraction task where binary relations are used to encode potentially novel processes/interactions between biomedical entities. Several existing BRE efforts focus on natural language processing (NLP) approaches that look at specific sentences for linguistic phenomena that are typically used to manifest/express different relations. In this effort, we demonstrate the potential of a complementary approach that employs graph patterns and instead models the prediction problem at a global level to output probability estimates of whether a pair of entities participates in a particular relation. Our models are trained with graph pattern features over the graph of predications from SemMedDB [1], a database of relations connecting UMLS entities with predicates, most of which are relation types, from the UMLS semantic network. We first introduced this approach [2] for the treats predicate and here we report improved results for both treats and causes predicates. Additionally, we report on the coverage provided by the top few patterns from the trained logistic regression (LR) models.

Methods: Given any two biomedical entities (as UMLS CUIs), our goal is to predict whether the entities participate in a treats or causes relation. For example, given Tamoxifen and Breast Cancer, our model for the treats predicate should output a positive decision as it is a well-known treatment relationship. The features we use are based on all simple paths of length ≤ 3 connecting both entities in the SemMedDB graph. We take each path and replace the nodes with their UMLS semantic type sets (given an entity can have more than one type) while keeping the predicate labels as they are. This essentially “lifts” specific paths to semantic patterns between the input entities. We call these patterns compound patterns [2] as each node is replaced with its semantic type set. A conventional LR model is trained with these patterns as features. In this effort, we also experimented with what we are terming “primitive” patterns where each compound pattern is expanded into multiple simpler patterns by replacing the semantic type sets with constituent individual items. Additionally, we also identified the coverage (recall) provided by the top few patterns when they are ranked based on LR model coefficients (which are log-odds ratios for the corresponding features).

We used 7000 treats and 2918 causes hand-curated relations from the UMLS Metathesaurus for our experiments and removed any existing treats and causes edges between the corresponding pairs from the SemMedDB graph when building the corresponding models (to ensure fair assessment of our approach). We derived negative examples from pairs that are not connected via the corresponding predicates in the Metathesaurus or the SemMedDB graph but are nevertheless allowed by the semantic network. We computed performance measures over 100 distinct 80%–20% train-test splits of the datasets to compute confidence intervals over F-scores with a positive-negative imbalance of 1:10 in the test set to model realistic scenarios. We varied the imbalance in the training datasets to see the corresponding effect on test set performance.

Results: For both treats and causes, models with compound pattern features outperformed their counterparts trained with primitive patterns although primitive patterns were superior when balanced training datasets were used. The training datasets with the most positive-negative skew (1:10) yielded the best performances in both cases. The 95% F-score confidence interval for the treats predicate is 98.03 ± 0.06. The corresponding interval for causes is 88.56 ± 1.8. For treats, we had over 423,000 unique compound pattern features in the model, but the top 100 patterns (based on model coefficients) cover 43% of the 7000 instances. Similarly, for causes we had over 197,000 unique compound patterns and the top 100 connect 25% of the full dataset. This shows that our method is also able to identify high-quality patterns that can be used to query knowledge graphs for generating new hypotheses.

References


Establishing Value of Health Information Exchange Content
Gina B. Baker, MSN1, Sidney N. Thornton, PhD1,2
1Intermountain Healthcare, Salt Lake City, UT; 2University of Utah, Salt Lake City, UT

Abstract
Health Information Exchange (HIE) has been proposed as a value added system to the expensive and complex healthcare problem of information silos and segmented patient treatment. Differing opinions of HIE success are circulating based on possible barriers, success stories and its promised potential. In an attempt toward objectivity, patient summaries within the CCDA format were evaluated for content and value added with respect to the current encounter. The ultimate usefulness of data can be defined by its inherent content and by the situational context to a provider at the point affecting care.

Introduction
Despite millions of dollars invested into health information exchange technology, questions of value and usefulness remain in regards to HIE systems. In a recent review Eden et al.1 identified the three main barriers for HIE as information completeness, organization, and workflow technology matching user needs. Beyond connectivity, questions remain about the information actually available for providers through health information exchange and whether the information is relevant. Information contained in a health information exchange depends on the architecture of the health information exchange, consent structure (opt-in or opt-out) and participating providers1. Once information is actually available in the HIE the problem becomes one of filtering the data according to its value or usefulness for the provider in the specific situation. In a study by Kierkegaard, Kaushal and Vest providers reported that they cannot find any patient information about 30% of the time and really only find what they are seeking about 5% of the time2.

Methods
Patient summaries from the community were reviewed against current visits for an internal medicine clinic. Data were analyzed based on the length of the summary, number of sources, data sections and types populated, whether the data was recent within the past six months and data completeness. A clinician evaluated the data points in addition to the scheduled visit reason compared to available HIE data context to determine if available external reason would be valuable to the current encounter. The HIE data was then categorized by its potential to alter the course of treatment.

Results and Discussion
Similar to the study by Kierkegaard, Kaushal and Vest2 patient clinical information was found from community HIE sources for 62% of the internal medicine visits. Patient summaries were generally three pages in length with an average of three external clinical sources. Patient results of any type were found 90% of the time while results within the last six months were noted 60% of the time with 25% being new information from the last visit at the internal medicine facility. Encounter information occurred 70% of the time in patient summaries followed by laboratory at 40% and radiology notes at 33%. Treatment course changing results were seen in the external data for only 1% of the visits.

Conclusions
Content of HIE will continue to grow as participation grows. Giving providers a snapshot of data content found in the exchange can help with provider acceptance and the ultimate rate of usage. Once information is readily available, the technology and clinical workflows need to consider data presentation, filtering and usability to aid the provider based on the specific encounter, specialty and the individual patient. Additional evaluations of HIE data in current visit context by multiple clinicians is recommended.

References
Mobile Interactive Case Simulation Tools to Assist Healthcare Providers in Clinical Decisions

Monica Barbosu, MD, PhD¹, Margaret Demment, PhD¹, Scott McIntosh, PhD², Tamala David, PhD, MPA, MS,FNP³, Jose Perez-Ramos, MPH¹, Terry Doll¹, Beatrice Aladin, MPA⁴, Cheryl Smith, MD⁴, Timothy Dye, PhD⁵

¹ Clinical and Translational Science Institute, University of Rochester School of Medicine, Rochester, NY; ² Public Health Sciences, University of Rochester, Rochester, NY; ³ University of Rochester School of Nursing, Rochester NY, The College at Brockport Department of Nursing, Brockport, NY; ⁴ New York State Department of Health AIDS Institute, New York, NY; ⁵ Obstetrics and Gynecology, Pediatrics, Public Health Sciences, Medical Informatics, Clinical and Translational Science Institute, University of Rochester, Rochester, NY

Introduction

The HIV/HCV/STD Clinical Education Initiative (CEI) provides continuing medical education to healthcare providers and is funded by the New York State Department of Health AIDS Institute (AI). In 2009 CEI diversified its portfolio with the addition of online education. To date we have offered more than 280 multimedia learning modules, over 100 online accredited courses, 14 interactive case simulation tools (ICST) and other informational materials to support continuing medical education among clinical providers. All these resources are delivered through various platforms: website (www.ceitraining.org), Android and iOS apps, social media and email newsletters. In 2013 we started offering interactive case simulations to healthcare providers as a fast and easy access tool to decision diagrams based on the latest AI guidelines.

The objectives of this paper are to describe the development of ICST, analyze the usage of these tools, and evaluate the tool based on the feedback from our users.

Method

The time period of this study is June 2013 to February 2016. The analysis from google analytics data on demographics and geographical distribution of users, total/new/returning users, mobile devices used, traffic sources and network referral indicated the ICS usage. Descriptive statistics and time series were used to examine all these variables and, whenever appropriate, correlations were also determined. To determine the merit of ICST, we considered the feedback received from users through our online survey.

Results

Simulation tools have been used by 17,345 users; 36% were returning users, 73% were in the 18-45 year age group, and 58% were female. Users accessed the simulation tools primarily from mobile devices (85%) compared to only 15% on computers. While targeted to New York State providers only half of the users (52%) came from the US; other countries included Tanzania (6%), India (5%), Kenya (4%), South Africa (3%), Brazil (2%), United Kingdom (2%), Ghana (1%), and Canada (1%). The most popular simulation tools were the HIV testing and PrEP. The interactive decision diagrams inside the tool are designed to offer step by step approach in different case scenarios.

The simulation tools were found to be easy to use and helpful in the process of medical decision making, providing a convenient and useful way to offer clinical education.

Conclusions

All ICST were developed based on the NYSDOH AI guidelines, using a new technology-enhanced simulation training approach to deliver up-to-date medical knowledge to healthcare providers. The survey data revealed that the majority of clinicians found the ICST easy to navigate and very useful in the process of medical decision. In addition, the tool is being used around the globe, especially in countries impacted by the HIV epidemic.
Constructing Diagnostic CT Exam Lists for Sites across an HIE

Anton O. Beitia, MD¹, Tina Lowry MS¹, Daniel J. Vreeman, PT, DPT, MSc², Bradley N. Delman, MD¹, Benjamin H. Slovis, MD¹, George Loo, DPH¹, Frederick Thum III, MD¹, Jason S. Shapiro, MD¹
¹Icahn School of Medicine at Mount Sinai, New York, NY; ²Regenstrief Institute, Indiana University School of Medicine, Indianapolis, IN; ³Department of Biomedical Informatics, Columbia University, New York, NY

Introduction

Duplicate computed tomography (CT) imaging may cause increased health care costs and unnecessary exposure to ionizing radiation. Since duplicate CT imaging often occurs because providers lack prior patient health information from other clinical sites, a prior CT alerting system based on results from a health information exchange (HIE) may reduce avoidable CT imaging. Such an alerting system requires decision rules that operate against a list of CT exam codes and corresponding descriptions across HIE sites. Radiology master files typically mix codes for all modalities within the same file, and also contain various administrative and billing codes. Since diagnostic CTs, rather than procedural/interventional CTs, are the most likely duplicate exam types, we propose a method for constructing lists of diagnostic CT exams from facilities within an HIE in order to focus resources on the most pertinent exam types.

Methods

Our overall approach was to query HIE server data with word stems to identify potential CTs and then apply filters to create separate lists of diagnostic, procedural and post-processing/administrative CTs. Radiology data from exam description fields contained on local HIE “edge” servers of sites participating in Healthix (New York City area HIE) were queried with strings likely to indicate CT exams (e.g., CT, CAT). For our study, we created three sets of terms that each contained all exams from 10 sites: a training set, a preliminary test set, and a final test set. For the training set, a user with domain knowledge extracted all non-CT, procedural CT, and administrative/post-processing CT exam descriptions and segregated them into separate tables. Following the removal of these non-diagnostic exams, by exclusion the CTs remaining on the original list were diagnostic studies. The lists of non-CT, procedural, and administrative/post-processing CT exam descriptions were analyzed to determine high occurrence strings that could serve as tokens for filtering exams into the three categories. Examples of strings used for filtering include laet and oetreo for non-CTs, biopsy and guidance for procedural CTs, and multiplanar, recon, and consult for post-processing and administrative CTs. We then applied these filters to the preliminary test set. One of the authors (AOB), a board certified radiologist with domain expertise, manually reviewed the resulting diagnostic CT list for false negatives (i.e., non-CTs, procedural CTs, and administrative/post-processing CTs that were included in the diagnostic list) and analyzed the non-CT, procedural, and administrative/post-processing CT lists for false positives (i.e., diagnostic CTs incorrectly excluded). Preliminary test set results were used to further refine the word stem filters, which were then applied to a final test set of potential CTs from 10 other sites. We calculated sensitivity and specificity of the filters for non-CTs, procedural, and administrative/post-processing CTs applied to the final test set.

Results

A total of 11,115 potential CT exam descriptions from 30 Healthix sites were obtained for the training, preliminary test, and final test sets. In the training set there were 3,034 potential CT exams of which 1,758, 777, 471, and 28 were diagnostic, non-CTs, procedural, and administrative/post-processing CTs respectively. In the preliminary test set there were 3,663 potential CT exams of which 2,357, 864, 392, and 50 were diagnostic, non-CTs, procedural, and administrative/post-processing CTs respectively. For the final test set, there were 4,417 potential CT exams of which 3,017, 1021, 333, and 46 were diagnostic, non-CTs, procedural, and post-processing/administrative CTs respectively. Sensitivity of the non-CTs, procedural CTs, and post-processing/administrative CTs refined filters applied to final test set were 0.98, 0.99, and 0.91 respectively. Specificity of these filters were 0.99, 1.00, and 0.97.

Conclusion

Our method of constructing CT exam lists from HIE sites produced diagnostic, procedural, and administrative/post-processing CT exam lists with a high sensitivity and specificity. Application of these filter methods may facilitate constructing HIE-wide and site-specific exam lists which can be used for terminology mapping, crossover analysis, and as the basis for a previous exam alerting system.
Categorizing Clinical Data to Make It Easier for Patients to Indicate Their Data Sharing Preferences

Elizabeth Bell, MPH, Diana Guijarro, Hyeon-eui Kim, RN, MPH, PhD, Alexander Richardson, Jina Huh, PhD, Shuang Wang, PhD, Lucila Ohno-Machado, MD, MBA, PhD
Health System Department of Biomedical Informatics, UC San Diego, La Jolla, CA

Introduction and Background
The informed CONsent for Clinical data and biosample Use for Research (iCONCUR) project aims at (1) studying patient data sharing preferences and (2) developing innovative web-based tools to facilitate the informed consent process. We previously developed a web-based tool prototype, iCONCUR v.1, which allows patients to make choices about which parts of their medical record they want to share with different categories of researchers. The choices are saved in a database that is checked before data are released to researchers. The choice taxonomy available to patients was based on interviews with healthy volunteers and analyses of the clinical data requests from the researchers in our institution. Our goal for the final system is for the data sharing options to cover the data items most commonly used in research, and for each item to be properly placed under an intuitive category that is easily understood by patients. To inform the taxonomy design for the final system including the categorization of data sharing options intuitive to the patients, we are conducting 4 focus groups, each consisting of 3~5 participants. This abstract reports on the process of developing data categories and results of the first focus group with 3 HIV+ patients.

Methods
To identify the medical data items to review in the focus group session, we examined 22 reference articles about informed consent and 192 nursing and clinical trial study papers. We identified more than 1,800 data items from these articles, which were later condensed to 1,100 unique items. Two readers collaboratively reviewed the 1,100 items and identified additional data items and categories that could potentially be included in the iCONCUR system. These articles helped us determine categories of clinical data that should be considered for the data sharing options. The information obtained from the articles was compared against the iCONCUR data to create a condensed version with 17 category labels (Table 1) to be used in the focus groups. The three participants of the first focus group performed a card sorting activity, sorting 67 medical data items into groups based on perceived similarities. After sorting all cards, they constructed category names for each group, informed by our 17 category labels. Letting the participants first sort the items and add category names later (as opposed to being presented with category names to sort the items into) allowed us to start to understand how some participants naturally group data items.

Results and Discussion
The participants developed 6 groups and sorted the items in unexpected ways. For instance, Neighborhood and Number of Pregnancies were grouped together under the category Family History, which these participants perceived as a category that would encompass any data item relating to their families. Additionally, patients did not place data items relating to sensitive topics such as drug use, mental health, or sexual history in a separate category. They also had a hard time considering more than two items in a group—i.e., in choosing which group an item should be included, these patients looked in each group for an item that was similar to the pending item, rather than considering multiple items in the group. These “anchor” items seemed to be those that the participants remembered more clearly because either they were easy to understand (e.g., height) or they had experience with it (e.g., urine sample). In future work, we plan to investigate how different card sorting or alternative procedures might have an impact on patients’ decisions. More focus groups with a larger number of patients from different clinics will give us insights to patients’ mental models on personal medical data categories. This will guide the development of iCONCUR v2.

Acknowledgements
This project is supported by NIH grants from NHGRI (R01HG008802 and R00HG008175).

Table 1. Seventeen categories for clinical data items (derived from the literature and used in iCONCUR v.1)

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Sub-categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family History</td>
<td>Medical History</td>
</tr>
<tr>
<td>Living Arrangement</td>
<td>Sexual History</td>
</tr>
<tr>
<td>Vitals</td>
<td>Anthropometrics</td>
</tr>
<tr>
<td>Behavioral History</td>
<td>Laboratory Tests</td>
</tr>
<tr>
<td>Demographics</td>
<td>Sociobehavioral</td>
</tr>
<tr>
<td>Health Encounter</td>
<td>Biological Samples</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Sub-categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropometrics</td>
<td></td>
</tr>
<tr>
<td>Sociobehavioral</td>
<td></td>
</tr>
<tr>
<td>Biological Samples</td>
<td></td>
</tr>
</tbody>
</table>
NASA’s GeneLab: Phase I Results and Plans

Daniel C. Berrios, MD MPH PhD\textsuperscript{1,2}, Jon D. Welch, BA\textsuperscript{2}, Homer W. Fogle, MS\textsuperscript{2}, Michael Skidmore, BS\textsuperscript{2}, Oana Marcu, PhD\textsuperscript{2}

\textsuperscript{1}University of California, Santa Cruz, CA; \textsuperscript{2}NASA Ames Res. Ctr., Moffett Field, CA

**Purpose**

GeneLab is currently being developed by NASA to support “open science” biomedical research in order to enable the human exploration of space and improve life on earth.\textsuperscript{1} Phase I of the four-phase GeneLab Data Systems (GLDS) project emphasized capabilities for submission, curation, search, and retrieval of genomics, transcriptomics and proteomics data from biomedical research of space environments. Key challenges included: 1) selecting and tailoring the platform for Phase I; 2) designing optimal data set curation procedures including metadata representation and generation, and quality control); and 3) balancing GLDS accessibility with user engagement and tracking capabilities.

**Design**

For GLDS Phase I, we deployed a project-management and file-sharing software platform developed by NASA, deemed the most expedient path towards democratizing data access to space life science research data. For metadata representation, the project surveyed a wide variety of options, and chose the ISA-Tab specification\textsuperscript{2} for its broad applicability to all kinds of biological experimental data, conceptual and practicable use of formal knowledge representation structures (i.e., ontologies), and availability of open-source metadata acquisition and translational tools.

**Results**

Open access to Phase I of the GeneLab Data Systems was implemented in April 2015. Usage of the system has steadily grown (see Figures 1, 2), mirroring the growth in curated data file collections (from 22 to 47 data sets as of Feb. 2016). Growth was observed despite limited end-user capabilities of the Phase I system (rudimentary data set browsing and search by metadata, limited metadata display). The majority of Phase I-curated data were transcriptomics investigations already housed in other repositories (a validation of data acquisition, storage and curation processes); four never-before-released, multi-omics spaceflight investigations (using Mus musculus and Arabidopsis thaliana), were added in late 2015, leading to an exponential increase in download volumes (Fig. 2).

**Conclusions**

The Phase I system download volumes indicate strong interest of the scientific community in these data. Phase II of the GLDS will focus on enhancing search capabilities, including search of federated repositories like the NCBI GEO/SRA and EBI PRIDE archives. We anticipate challenges in the areas of federation flexibility, federated search performance, and metadata representation mapping. More powerful search, data and metadata visualization functions, and more automated data submission processes may increase usage dramatically. The project is currently evaluating how best to include these capabilities, including evaluating alternative software platforms for Phases II-IV.

**References**

Toward Honoring the Values of Patients With Multiple Chronic Conditions: Insights from a field study

Andrew B. L. Berry, MS¹, Catherine Lim, MDes¹, Tad Hirsch, PhD¹, Andrea L. Hartzler, PhD², Edward H. Wagner, MD, MPH², Evette Ludman, PhD¹,², James D. Ralston, MD, MPH¹,²
¹University of Washington, Seattle, WA; ²Group Health Research Institute, Seattle, WA

Introduction
Older adults with multiple chronic conditions (MCC) face competing demands for care, which can impede self-care among patients and their family caregivers¹. Furthermore, patients and their healthcare providers often disagree on priorities for care—providers tend to focus on managing individual conditions while patients often focus on their ability to engage in meaningful activities². It is important that patients and providers set care priorities together. Patients’ values—things that are important to them in the context of their daily lives—influence their priorities, but previous work has defined patients’ values inconsistently³. Therefore, we sought to clarify and expand the definition of patients’ values with the ultimate goal of understanding how their values affect their priorities for care.

Methods
We conducted 24 home visits with patients who receive care at a large integrated healthcare system in Washington state. Home visits involved a semi-structured interview, aided by photo elicitation, and a home tour. The goal was to understand what is important to patients in their daily lives, and how they relate what they value to their illnesses and self-care. We recruited adults with diabetes and at least two of the following: osteoarthritis, depression, and coronary artery disease. Our analysis of interview transcripts was guided by constructivist grounded theory.

Results
Five themes about what patients value emerged (Table 1). Participants often described relationships among values, such as tensions and synergies. For example, P2 expressed tension between her volunteer work providing food for the homeless and her desire to avoid pain and fatigue: “The first thing I’m going to have to give up is, unfortunately...I hate to even verbalize it...is [volunteering]. That’s the most strenuous part, and the most risky, that [as] an 85-year-old may not be the best position for me to be in...when I go to Costco, and lifting all that stuff in the car, my shoulders, pain, you know?” P2’s volunteer work represents a valued activity, but also reflected other values, including principles, emotions, and social relationships. To continue volunteering, P2 scheduled days off to rest and read in her recliner, reflecting her value of comfort. Findings illustrate different things patients value and suggest that clinicians should strive to understand relationships among patients’ values and demands for care.

Conclusion
We characterize what patients with MCC value and relationships they perceive among those values. Findings establish a foundation for future work to understand how patients’ values shape their care priorities, how patients and providers communicate about values, and how to design systems to promote shared priorities for care.

Acknowledgments: Study procedures were approved by Group Health Research Institute IRB. We wish to thank the Agency for Healthcare Research and Quality for their generous support (R01 HS022364) and valuable contributions of Linda Kiel, Zoe Bermet, and Luesa Jordan.

References
Transmission of ELR messages To Improve Public Health Reporting

Samar Binkheder, MS, PhD Cand1, Brain E. Dixon, PhD2,3, Shaun Grannis, MD, MS3,4
1Indiana University School of Informatics and Computing, Indianapolis, IN; 2Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; 3Regenstrief Institute Center for Biomedical Informatics, Indianapolis, IN; 4Indiana University School of Medicine, Indianapolis, IN

Introduction

The notifiable condition detector (NCD) receives more than 350,000 Health level 7 (HL7) messages daily for public health reporting (1). The NCD system showed the lowest sensitivity 38.0% for free-text (Methicillin Resistant Staphylococcus Aureus (MRSA)), 70.2% for discrete (Influenza), and 97.1% for numeric (Lead) (2).

Methods

First, we manually evaluated a previously extracted set of 100 HL7 microbiology messages for their compliance with HL7 rules, and patterns of errors of electronic laboratory reporting (ELR). Second, we performed a descriptive analysis of the received HL7 messages (including non-microbiology test results) for one month (August 2015) for percentages of microbiology results, LOINC, and SNOMED. Text processing and descriptive analysis were performed using Python language regular expressions and SAS.

Results

The analysis revealed several patterns. For example, free-text results (e.g. MRSA) were written in multiple ways for 22% of messages. In addition, multiple results or organisms were written as one result in HL7. Moreover, even though some of the results were in numeric format, they were reported as a text format. Additionally, some irrelevant information was found within the result, such as address or person’s name.

During August, 2015 the NCD has successfully identified a total of 31,831 (excluding neoplasms) notifiable conditions. We found that microbiology-related totally 8,656, which is 27.2% of the notifiable conditions. “Staphylococcus, Methicillin-resistant” was the most prevalent (N=2950; 9.3%) microbiology-related condition. In order to learn about the reportable conditions in free-text format, LOINC codes were extracted only for the reported free-text format results. There were a total of 150,622 LOINC codes (27.2%) represented in the data set. For all of the one-month income messages, adoption of SNOMED CT codes was low with only 1323 codes (1%).

Conclusion

This work is still in progress. Even though the NCD has successfully identified notifiable conditions from text data, it is still important to analyze the non-reported conditions to increase the sensitivity and specificity of detection. Results from this work will identify message problems and compliance with standards. In addition, our results will facilitate the identification of methods that improve transmission of ELR messages, and increase the use of structured data. Future work will also seek to apply machine learning and natural language processes techniques that have proven useful in analyzing text documents in other biomedical informatics application spaces.

References

Comparison of Approaches for Heart Failure Case Identification from EHR Data

Saul Blecker, MD, MHS1, Stuart D. Katz, MD, MS1, Leora I. Horwitz, MD, MHS1, Gilad Kuperman, MD, PhD2, Hannah Park, MS1, David Sontag, PhD3
1NYU School of Medicine, New York, NY; 2New York-Presbyterian Hospital, New York, NY; 3New York University, New York, NY

Background: Heart failure (HF) is a common cause of hospitalization and a focus of readmission reduction efforts. Accurate, real-time case identification is needed to target interventions to improve quality and outcomes for hospitalized patients with HF. Problem lists may be useful for case identification, but are often incomplete. Machine learning approaches may improve accuracy of case identification, but may be limited by complexity. We developed five algorithms of increasing computational complexity to identify hospitalized HF cases.

Methods: We performed a retrospective study of hospitalizations at a single academic medical center in 2013-2015 using electronic health record (EHR) data. Using a development set comprised of a random sample of 75% of hospitalizations, we developed 5 algorithms for HF identification at the second midnight of hospitalization: 1) HF on problem list; 2) presence of at least one of three characteristics: HF on problem list, any inpatient diuretic use, or brain natriuretic peptide≥500 pg/ml; 3) logistic regression of 30 clinically relevant structured data elements; 4) machine learning approach using L1-regularization logistic regression with unstructured notes and imaging reports; 5) machine learning approach using both unstructured data and 30 clinically relevant structured data elements. For algorithms 3, 4, and 5, we calculated areas under the receiver operating curves (AUCs) as well as sensitivity based on setting a positive predictive value (PPV) at 0.9. Model validation was based on physician review of selected charts from the 25% validation set.

Results: Of 47,119 hospitalizations, 6,549 (13.9%) had a discharge diagnosis of HF. Inclusion of HF on the problem list (algorithm 1) had a sensitivity of 0.40 and PPV of 0.96 for HF identification in validation. Algorithm 2 improved sensitivity to 0.77 at the expense of PPV of 0.64. Algorithms 3, 4, and 5 had AUCs of 0.953, 0.969, and 0.974, respectively (Figure). With PPV of 0.9, these algorithms had associated sensitivities of 0.68, 0.77, and 0.83, respectively.

Conclusion: The problem list is insufficient for real time accurate identification of HF hospitalization cases. Although analysis of free text using machine learning had the highest predictive accuracy, less computationally-intensive approaches using structured data may be adequate depending on clinical need.

Figure. Receiver operating curves (ROCs) for three algorithms to classify patients with heart failure: logistic regression of structured data (algorithm 3), machine learning of unstructured data (algorithm 4), and machine learning of a combination of structured and unstructured data (algorithm 5).
Characterizing the semantic composition of the UMLS Metathesaurus over time

Olivier Bodenreider, M.D., PhD and Lee Peters, M.S.
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
{olivier.bodenreider|lee.peters}@nih.gov

Motivation. The UMLS Metathesaurus has grown dramatically over the past fifteen years. From 2002 to 2015, the number of concepts has increased from about 777,000 to 3.1 million, a 4-fold increase. It is difficult to infer the semantic composition of the UMLS from the list of its sources. While some source vocabularies contribute concepts from a single semantic category (e.g., anatomical entities in the Foundational Model of Anatomy), others reflect a wide range of semantic categories (e.g., SNOMED CT). Moreover, the integration of a given term from a source vocabulary does not always result in new concepts, since this term may end up as a synonym for an existing concept. In this investigation, we leverage the semantic groups to characterize the semantic composition of the UMLS Metathesaurus and its evolution over time.

Methods. Each Metathesaurus concept is assigned at least one of the 127 semantic types. Semantic types are grouped into fifteen semantic groups, which represent broad subdomains of biomedicine, such as Anatomy, Chemicals and Drugs, and Disorders. The UMLS semantic groups have been used to create semantic profiles for source vocabularies, but can also be applied to the Metathesaurus as a whole. For each edition of the UMLS (2002-2015), we compute the distribution of the Metathesaurus concepts with respect to the 15 semantic groups.

Findings. As shown in Figure 1, the proportion of a few semantic groups has changed markedly between 2002 and 2015. Chemicals & Drugs. The composition of the early Metathesaurus versions was heavily dominated by chemical concepts, especially from MeSH. Although the number of chemical concepts has almost doubled during this period, it has grown at a slower pace than that of other groups. Living Beings. This group mostly represents organisms (mainly from the NCBI taxonomy) and has grown from 30,000 to nearly 1M concepts. It is the fastest growing group and now represents 30% of all Metathesaurus concepts. Disorders and Procedures. The integration of a single large and fine-grained vocabulary can be responsible for major shifts in composition. The growth of disorder concepts between 2009 and 2011 is attributable to the integration of two large and fine-grained vocabularies, MEDCIN (2009) and ICD10-CM (2011). This is also the case for procedure concepts in 2009, when ICD10PCS was added. In contrast, the integration of SNOMED CT in 2003-2004 is silent, because, although extensive and detailed, its content was already largely represented by SNOMED International and the Read Codes.

For information about the UMLS semantic groups, see: https://semanticnetwork.nlm.nih.gov/.

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Figure 1. Evolution of the semantic group distribution in the UMLS Metathesaurus
Identifying Missing Finding Site Relations in SNOMED CT

Jonathan P. Bona, PhD\textsuperscript{1}, Werner Ceusters, MD\textsuperscript{1},
\textsuperscript{1}Department of Biomedical Informatics, University at Buffalo, Buffalo, NY

Introduction

Every concept in SNOMED CT comes with descriptions, including a Fully Specified Name (FSN). FSNs typically end with a semantic tag in parentheses to disambiguate from other concepts with similar descriptions. However, some things that are not semantic tags appear in parentheses as parts of FSNs, and some FSNs have more than one tag, so it’s sometimes hard to know by looking at an FSN exactly what its tags are. We will call words within parentheses in FSNs apparent tags if it’s not obvious that they are not tags. “(bullous edema is not sufficient evidence to classify a tumor as T4),” appears in an FSN but is clearly not a semantic tag.

SNOMED CT’s concept model includes associative relations between concepts used to formally specify their meanings. For example, 10000006 | Radiating chest pain (finding) is associated via 363698007 | Finding site (attribute) with 51185008 | Thoracic structure (body structure). Formally specified and query-able relations between concepts enhance the completeness and usefulness of an ontology, and increase its utility for organizing and retrieving data, e.g., as part of structured EHRs, especially when combined with the Is-a hierarchy and subjected to logical inference. SNOMED CT’s main finding concept 404684003 | Clinical finding (finding) has many thousands of sub-concepts. Here we examine the use and disuse of finding sites in finding concepts related to cancer staging (specifically, some of the 1038 concepts under 385356007 | Tumor stage finding (finding)). For those lacking finding sites, we use apparent tags in the FSNs to find body structure concepts that should be associated with those findings.

Method

We first found all Tumor stage finding concepts that lack finding site relations to any body structure, and identified 170 candidate semantic tags in their FSNs. These were manually filtered for likely body structure tags: apparent tags that name anatomical entities. We focused on the 47 simplest body structure tags and omitted tags with longer phrases (e.g., “liver, including intrahepatic bile ducts”), and 36 short “tags” like “pT1-pT4, pM1” that are codes for different components of tumor staging. Each of these 47 body structure tags was manually matched to the best fit body structure concept. Starting again from the full set of Tumor stage findings, we used the body structure tags and matching concepts to generate a list of suggested finding sites for any of the 1038 concepts that used those tags.

Results

The result is a list of suggested candidate finding sites for 369 concepts (out of 1038) that do not have associated finding sites. These were manually verified as coherent and plausible finding sites for the relevant concept. Each concept was then marked with its nearest anatomically-specific TNM finding ancestor concept (e.g. 397092003 | Gallbladder TNM finding (finding)) where applicable. Almost all had anatomically-specific TNM ancestors. Of the 17 concepts without such ancestors, 6 have the apparent tag “appendix.” Unlike other body structures (Bone, Thyroid, etc) there is no TNM finding concept for Appendix. 4 out of 17 are pN (regional lymph nodes) or pM (distant metastasis) breast cancer findings. Another was for a primary tumor that cannot be assessed. Since none of these is about things located in the breast, they might need no finding site relation with Breast (body structure). The prevalence of anatomically-specific TNM finding concepts here suggests another technique to efficiently add finding sites for many of the tumor stage concepts: such concepts (Ampulla of Vater TNM Finding, etc) should be directly associated with the relevant body structures (Structure of ampulla of ...) -- relations that will apply also to their descendants (pT1: Tumor limited to ampulla of Vater ... ) after classification.

Conclusion

Our preliminary investigation of the use of finding site associations in a small section of the SNOMED hierarchy reveals an important gap that limits the usefulness of the relevant concepts for representing and reasoning about cancer staging and its relation to patients, parts of their bodies, their health, and so on. We have used the presence of apparent semantic tags that name body structures to align approximately one third of the target concepts with body structure concepts suitable to serve as their finding sites.
Pre and Post Knowledge Self-Assessment in a Flipped Classroom: Analysis of Students in an Introduction to Health Informatics Course

Suzanne A. Boren, MHA, PhD\textsuperscript{1}, Iris Zachary, PhD, MS, CTR\textsuperscript{1}, \\
\textsuperscript{1}Department of Health Management and Informatics, University of Missouri, Columbia, MO

Abstract

Self-assessment of knowledge by students is an important tool for course design, learning, and assessment in a flipped classroom. The students in an introductory health informatics course completed knowledge self-assessment surveys covering the full content of the course at the beginning and end of the course. The self-assessment served as an outline of the learning objectives for the course, a resource for students to understand their learning, and a measure of their gains in knowledge.

Introduction

The purpose was to evaluate the pre and post knowledge of master of health administration and master of health informatics students enrolled in an introduction to health informatics course taught in residential and executive formats. Education models continue to move to student-centered learning. Adopting methods for understanding student learning must be utilized as part of a comprehensive assessment system. In addition, formative assessment and feedback should be used to empower students as self-regulated learners.

Method

Students were invited to complete a self-assessment exercise of their health informatics knowledge at the beginning and end of the course. Sixty-eight (68) students were enrolled in the course and participated in the pre and post knowledge self-assessment. One hundred thirty-four (134) knowledge areas were selected from the chapters of an introductory health informatics textbook. The students rated each knowledge area on a five point scale with the corresponding anchors: no knowledge (1), very limited knowledge (2), limited knowledge (3), moderate knowledge (4), great knowledge (5). The data was collected, de-identified, organized, and analyzed to calculate the change in knowledge between the beginning and end of the course. The pre and post knowledge self-assessment means were calculated for each student to evaluate how each individual benefited, as well as the class as a whole. Following this, the pre and post knowledge self-assessment means were calculated for each chapter. Each of the 134 areas was also linked to a level of Bloom’s Taxonomy.

Results

Overall, there was a pre knowledge mean of 2.0 (limited knowledge) and a post knowledge mean of 4.3 (moderate knowledge). The pre assessment overall mean for each student ranged from 1.0 to 3.3. The post assessment overall mean for each student ranged from 2.8 to 5.0. Four of the 68 students reported post assessment knowledge scores of 5.0 indicating great knowledge in all areas. The overall difference between pre and post knowledge self-assessments ranged from 0.5 to 3.7. The pre knowledge overall mean for each chapter was an average of 2.0 and the post knowledge overall mean was an average 4.0.

Conclusion

The knowledge gained during the semester of an introductory health informatics course was quite satisfactory and moved 2.3 points on a 5.0 point scale from limited knowledge to moderate knowledge. While a pre and post knowledge self-assessment surveys cannot be used solely to determine a student’s learning, they can certainly be used as part of a comprehensive assessment plan for validation of learning. Implications for educational practice include a greater awareness by students and instructor of course content requiring additional emphasis.
Evaluation of Safety Heuristics Applied to a Mobile Medication Reconciliation Application

Elizabeth M. Borycki, RN PhD1, Andre W. Kushniruk, MSc PhD FACMI1
1University of Victoria, Victoria, British Columbia, Canada

Abstract

There is a need to evaluate the ability of safety heuristics to identify potential technology-induced errors in mobile health (mhealth) applications. In this work we evaluate the ability of safety heuristics to identify potential technology-induced errors in mhealth applications.

Introduction and Background

The safety of health information technology (HIT) has emerged as an important issue for the modern healthcare organization. New technologies are being developed that can improve the quality and safety of health care, yet, along with the introduction of these new HIT there has also emerged new types of errors (i.e. technology-induced errors) 1. Researchers have developed a number of differing methods for identifying technology-induced errors among them safety heuristics2. Safety heuristics have been used to evaluate software applications such as EHRs that are accessible via desktop and laptop computers. Less research has been done on the assessing the safety of mhealth applications using safety heuristics2,3. In this poster we describe our pilot work in applying safety heuristics developed for traditional EHRs available via desktop and laptop computers to mhealth applications.

Methods

There exist a number of mhealth applications that health professionals and consumers have identified as being supportive of patient safety. One type of application that has been identified by consumers and safety organizations as being of value is the mobile medication reconciliation application (mMedRec app)4. Many safety organizations recommend using mMedRec apps to help patients remember the medications they are taking, remind them to take the right medication, the right dose of the medication and at the right time. To assess the ability of published safety heuristics to evaluate the safety of an mhealth MedRec application, an analyst stepped through the user interface of a MedRec app using safety heuristics developed by Carvalho et al for EHRs2,3.

Results

It was found that many of the safety heuristics developed by Carvalho et al2,3 could be used to evaluate an mMedRec app. The safety heuristics are subdivided into four safety areas: workflow, content, safeguards and function. Workflow, content and safeguard heuristics helped to identify mhealth application safety issues (i.e. those associated with user interface design and workflow)2,3. Several “function” heuristics focused on interoperability and integration of mhealth applications with EHR software2,3. The application studied indicated problems regarding integration and exchange of data with EHRs. This work suggests that interoperability and exchange of data between mhealth apps and EHRs is a key aspect of mhealth and patient safety especially for mMedRec applications.

Conclusion

The work identifies that there is a need for heuristic evaluation of mMedRec apps. In particular there is a need for further study of the integration and exchange of data between consumer mMedRec apps and traditional EHRs and their implications for patient and technology safety.

References

Opioid-Related Side Effects and Adverse Reactions: Mining of Patient Secure Messages

Lina Bouayad, PhD1,6, Robert D. Kerns, PhD2, William Becker, MD2,3, Thomas Houston, MD, MPH4, Stephen L. Luther, PhD1, Stephanie Shimada, PhD4,5, Balaji Padmanabhan, PhD6

1James A. Haley Veterans’ Hospital, Tampa, FL; 2Yale University, New Haven, CT; 3VA Connecticut Healthcare System, New Haven, CT; 4Bedford VA Medical Center, Bedford, MA, 5Boston University, Boston, MA; 6University of South Florida, Tampa, FL

Introduction

Extensive data about opioid-related side effects and risk factors to predict opioid-associated adverse events are stored in the EHR in coded fields and progress notes. Research indicates the structured data in EHR are incomplete and inconsistent. Prior literature has targeted mining of clinical documents written by providers. Yet, Patients’ data extracted from unstructured progress notes contains provider-reported information only. Recently, secure messaging has been developed as a tool that provides patients with convenient and fast access to healthcare providers regarding non-urgent health concerns. As of October 2015, an exponential growth in SM adoption was reported among veterans with more than over 1 million SMs sent. This suggests that SMs represent an important new resource that was previously inaccessible for clinical care or research.

Patient-reported side effects and adverse events can be extracted from secure messages (SMs). Data extracted from SMs could be integrated into the EHR (as structured fields) to (1) enhance predictive risk models for opioid-related adverse events, and (2) improve time of responses by health care providers for patients with pain.

To our knowledge, this would be the first study to use analytics to explore the content of SMs. If successful, the method could be applied to other high-frequency, high-risk, and/or high-cost conditions.

Methods

This retrospective study looked at cohort of Veterans who 1) have been diagnosed with pain, 2) have been prescribed opioid therapy during FY12-FY13, and 3) have used SM to communicate with clinicians. Analyzing our sample data, we identified 1,300,885 unique Veterans who were prescribed common opioids (hydrocodone, oxycodone, morphine, codeine, tramadol, methadone, and hydromorphone) between FY12 and FY13. We then searched for keywords indicating opioid-related side effects and adverse events within the secure messages associated with that cohort of patients (exceeding 2 million). Then, Machine Learning was applied to identify the different topics documented in the selected secure messages. Topics identified through machine learning will be presented at the conference.

Results

Using secure messages’ data (through the VA Informatics and Computing Infrastructure (VINCI)), we able to complete two preliminary investigations. Results indicating the prevalence of keywords indicating opioid-related side effects and adverse events are listed in Table 1. Results also indicate that terms used by patients to describe health concerns in SMs can be different than the ones used by clinical providers. Therefore, future text mining work on SMs needed to be adapted to include commonly used patients’ terms. Co-occurrences of symptoms identified through data mining will be presented at the conference.

Conclusion

In the present study we assessed the presence of opioid-related side effects and adverse events in patients’ secure messages. Our future work intends to use text mining and machine learning algorithms to extract symptoms related to opioid therapy from SMs which will enable (1) comparison between patient- and provider-reported data and (2) integration of patient-reported symptoms into structured fields for further analysis.

References


<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain</td>
<td>469,273</td>
</tr>
<tr>
<td>Tired</td>
<td>40,067</td>
</tr>
<tr>
<td>Fall</td>
<td>39,719</td>
</tr>
<tr>
<td>Headache</td>
<td>36,823</td>
</tr>
<tr>
<td>Migraine</td>
<td>16,525</td>
</tr>
<tr>
<td>Fever</td>
<td>16,176</td>
</tr>
<tr>
<td>Nausea</td>
<td>15,797</td>
</tr>
<tr>
<td>Dizziness</td>
<td>10,885</td>
</tr>
<tr>
<td>Sweat</td>
<td>10,479</td>
</tr>
<tr>
<td>Energy</td>
<td>9,906</td>
</tr>
<tr>
<td>Fatigue</td>
<td>9,661</td>
</tr>
<tr>
<td>Confusion</td>
<td>9,576</td>
</tr>
<tr>
<td>Vomit</td>
<td>8,822</td>
</tr>
<tr>
<td>Constipation</td>
<td>8,723</td>
</tr>
<tr>
<td>Lightheaded</td>
<td>2,639</td>
</tr>
<tr>
<td>Dry mouth</td>
<td>2,415</td>
</tr>
<tr>
<td>Overdose</td>
<td>1,073</td>
</tr>
</tbody>
</table>

Table 1: Term SM Frequency
Home3D -- Virtualized Home Environments in the EHR
Markus Broecker, PhD, Patricia Flatley Brennan, PhD, RN
Living Environments Lab, University of Wisconsin-Madison

Problem
Patients’ homes play an increasingly important role in health care, from being a place for self-care and disease management to a source of health hazards or barriers and facilitators of living healthy. However, there is little space and no formalized way to capture in the electronic health record the typical ways clinicians learn about homes: verbal reports from the patient, still photographs, videos, or home visits. It is difficult to gain a shared understanding by listening to verbal reports or viewing images, and it is cost-prohibitive to make visits to the home of every patient who might benefit from this. Our team has developed a way to capture high fidelity 3D images of home interiors and display them in immersive virtual reality (VR) platforms in a way that enables rich understanding and sharing of perspective. Yet to make these visualizations clinically useful it is critical to find a way to integrate them into the workflow and generate entries in the EHR of the home images and of the decisions based on them.

We have partnered with a local medical center (UW Health) and a major electronic health records vendor (Epic) to demonstrate how to access and explore full 3D images of a patient’s home, annotate the images, and store them in a clinical record. Other members of our team are adapting existing standard terminologies to label specific objects in the home. During this one-year proof of concept project we will 1) refine a web-based application that enables omni-perspective viewing of point cloud data; 2) integrate this web viewer application into the Epic clinical interface; and 3) evaluate through focus groups of patients and clinicians the relevance and practicality of being able to access and explore a full scale model of a patient’s home. This strategy will be useful for those patients with complex self management needs, those at risk for household hazards, and those with discharge planning needs.

Approach
Figure 1 depicts the workflow. In step A, a point cloud of the home is generated and processed (currently using a $40,000 LiDAR scanner; future work anticipates generating point clouds simply and cheaply via smartphone-captured video). Because of the very large file size (> 1Gb) of the point cloud the storage in Step E must be external to the EHR using existing DICOM standard. Our project focuses on the clinical informatics challenges noted in Steps B, C, and D. At the appropriate time after a point cloud of a patient’s home is available, the clinician and patient can review the patient’s anticipated home patterns using any 3D visualization platform (e.g. VR head-mounted display, such as the Vive). Together they can make annotations of relevant spaces, for example where dressing changes can be done easily or where medication can be stored safely. Screen-scraped annotated images are created (C) and only those images sent to the electronic health record (D).

Current Status
Using HTML5 and WebGL standards we developed prototype web viewer application, which will be accessed via ‘tab’ from the EHR screen displayed on any platform or through immersive VR platforms. The existing data pipeline will be optimized to create data optimized for server-side storage and asynchronous file transfer. The current version of the code is available from github (https://github.com/broecker/PointCloud-WebGL).

Anticipated Challenges
We anticipate two key issues: 1) technological challenges, such as guaranteeing interoperability with existing data types and implementation details; and 2) integration into the existing clinical workflow, such as training to personnel to use the viewer and analyze the data, determining the correct patient population for which such information might prove useful or providing guidelines addressing privacy concerns of patients.

References

Supported by a grant from the AHRQ R03 HS 02462
A Systematic Review of The Types And Causes Of Prescribing Errors Generated From Using Computerized Provider Order Entry Systems in Primary and Secondary Care

Clare L. Brown, MPharm, Helen L. Mulcaster, Katherine L. Triffitt, BSc, Dean F. Sittig, PhD, Joan Ash, PhD, Katie Reygate, MPharm, Andrew K. Husband, PhD, David W. Bates, MD MSc, Sarah P. Slight, PhD

1School of Medicine, Pharmacy and Health, Durham University, U.K.; 2Newcastle upon Tyne hospitals NHS Foundation Trust, U.K.; 3The University of Texas Health Science Center, School of Biomedical Informatics at Houston, TX, USA; 4Oregon Health & Science University, Portland, OR, USA; 5Health Education KSS Pharmacy, U.K.; 6The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, USA; 7Harvard Medical School, Boston, MA, USA; 8Harvard School of Public Health, Boston, MA, USA.

Abstract: Computerized provider order entry (CPOE) systems have been shown to reduce the occurrence of prescribing errors. However, reports have also emerged that these systems have contributed to new types of errors. We conducted a systematic review of the literature to understand the different types and causes of prescribing errors associated with CPOE systems. We identified eight key themes in which improvements could be made.

Introduction: CPOE systems with clinical decision support (CDS) have been shown to reduce the occurrence of prescribing errors. However, a study by Koppel et al. in 2005 sought to identify and quantify the role of CPOE in facilitating prescription errors. Since then, many more studies have provided rich insights into the types and causes of these CPOE related prescribing errors. We conducted a systematic review of the literature to understand the different types and causes of these errors, and recommend improvements in these systems.

Methods: Our review was conducted according to PRISMA guidelines. We searched the literature published between January 2004 and June 2015 using three large databases: the Cumulative Index to Nursing and Allied Health Literature (CINAHL), Embase and Medline, and terms relating to ‘CPOE’ and ‘Errors’. Three authors reviewed titles and two authors reviewed all abstracts and full texts. Studies that reported qualitative data about the types and causes of these errors were included. Three authors read and re-read all full text articles and key recurring themes and sub-themes were identified iteratively from the data.

Results: We identified 1,185 publications, of which 34 were included. We identified eight key themes related to the causes of different types of prescribing errors associated with CPOE systems: computer screen display, drop-down menus and auto-population, wording, default settings, non-intuitive or inflexible ordering, repeat prescriptions and automated processes, users’ work processes and alerting. For instance, the layout of the computer screen display affected how users viewed patient information. Displaying an incomplete list of a patient’s medications on the computer screen contributed to a drug overdose being prescribed for a patient. Intravenous medications may not be displayed in the area of the screen where the patient’s other medications were, and thus users were less likely to consider these medicines when prescribing. Lack of system flexibility also resulted in users employing error prone workarounds, such as the addition of contradictory free-text comments, and CDS alerts generated inappropriate prescribing recommendations. Misinterpretation of the terminology used to describe data labels was also linked with the occurrence of prescribing errors.

Conclusions: This review highlighted how there is still much to be done to improve the safety of these systems. Human factors and user-centred design is key across all of these eight areas and should be prioritized by developers. Drop-down menus should be designed with safeguards to prevent selection errors. A crucial development will be the production of more patient specific and better worded alerts to reduce the impact of alert fatigue and erroneous suggestions.

An outcome-weighted network model for quantifying and measuring collaboration in a hospital cardiology unit

Matthew B. Carson, PhD1, Denise M. Scholtens, PhD1, Conor N. Frailey, PhD1, Stephanie J. Gravenor, MS1, Emilie S. Powell, MD, MS, MBA1, Amy Y. Wang, MD1, Gayle S. Kricke, MSW1, Faraz Ahmad, MD1, R. Kanna Mutharasan, MD1, Nicholas D. Soulakis, PhD1

1Northwestern University, Chicago, IL

Introduction: Shared patient encounters form the basis for collaborative relationships, which are crucial to the success of complex and interdisciplinary teamwork in healthcare. Quantifying the strength of these relationships using shared risk-adjusted patient outcomes provides insight into interactions that occur between healthcare providers in a hospital setting. A network-based approach to quantify teamwork quality could characterize clinical processes, facilitate quality improvement (QI), and become an important tool in learning healthcare systems.

Objective: To establish a generalizable, graph-based framework for calculating and measuring the Shared Positive Outcome Ratio (SPOR), an objective composite measure that quantifies the concentration of risk-adjusted positive outcomes for each pair of providers over a set of shared patient encounters.

Methods: We extracted patient encounter data from a cardiology unit over a three-year period from the Northwestern Medicine Enterprise Data Warehouse (NM EDW). For each encounter, we identified a list of associated healthcare providers along with an outcome indicating the patient’s satisfaction and an acuity measure that we used for risk-adjustment modeling of the encounter outcome. Next, we created a provider-encounter network and used this to identify providers who shared encounters. Finally, we evaluated the SPOR by constructing a network model to assess pairwise collaboration and compared each SPOR in the cardiology collaboration network to SPORs for corresponding edges in 1000 randomly permuted networks.

Results: From a network of 334 providers and 3,453 relationships, we identified 188 pairwise collaborations resulting in significantly high patient satisfaction rates (p ≤ 0.05). To examine high-scoring collaborative relationships more closely, we identified the providers who had 1) ≥ 10 collaborations involving ≥ 6 encounters, and 2) ≥ 10% of their total collaborative interactions with a significant SPOR. Twenty-six providers matched these criteria, indicating potential top performers in terms of patient satisfaction. Though providers in the high- and low-scoring groups had the same average number of associated encounters (41), those in the high-scoring group had a higher percentage of total encounters with positive outcomes than those in the low-scoring group (78% vs. 64%).

Conclusions: We identified extreme high- and low-scoring relationships over a set of shared patient encounters and quantified high variability in collaboration between providers. Our study shows that a healthcare collaboration network can be structurally evaluated to gain insight into the collaborative interactions that occur between healthcare providers in a hospital setting.

Figure 1. An example provider collaboration network showing 22 providers and 30 SPOR relationships. Properties associated with the highlighted edge (yellow) including the SPOR rank, an indication of the significance of the SPOR coefficient (pval), the SPOR coefficient, and the number of shared patient encounters between the two providers (num_collabs), are shown in the bottom left. The proximity of nodes to each other is based on the SPOR coefficient, with high-scoring relationships being shorter in length and low-scoring relationships being longer.
Extending Informatics Education to Home Care

Gail R. Casper RN, PhD¹, Paula A. Jarzemsky RN, MS¹, Jennifer D. Athanas RN, MS¹, Patricia Flatley Brennan, RN, PhD¹

School of Nursing, University of Wisconsin-Madison, USA

Background

Investments by ONC to develop the informatics workforce largely focus on preparing health professionals and informatics specialists to develop, deploy and evaluate information technologies within health care institutions that are familiar and readily accessible, in which their work responsibilities are well-known. As health care increasingly takes place in homes, lay people join professionals as essential health care workers; thus preparing informatics professionals for these less-formal, less accessible health care environments is needed. Patients and their families are required to manage unfamiliar health care demands and technologies in their homes. Critical aspects of the home such as the size, layout and condition are often unknown to those who prepare patients for discharge, prescribe requisite care or provide direct home care. Patients and families need guidance in planning and integrating their health care within their homes’ physical spaces; yet most health professional students lack curricular opportunities to master these skills. While the research on the effectiveness of simulation in clinical care is growing, these studies focus on clinical competencies such as communication, psychomotor skill acquisition and decision-making without risk in formal care settings¹. The home is a site of care, and simulated practice in the home has not been studied.

As part of the vizHOME project (www.vizhome.org; HS R01HS022548) our team captured and rendered full, 3D replicas of the interiors of 20 homes. We now use these replicas to enable health professional students to experience and understand the range and variability of such intimate personal spaces, and to enable them to generate guidelines about the location, use and expected benefits of computer and information technologies in the home.

Methods

Health professional trainees first experience a 3D home replica in a single point-of-view video exploration of an actual home with its characteristic “lived-in” look – a look that has not been successfully created in previous virtual “worlds.” In this “insitu” training, we guide the students to design, implement and evaluate a variety of simulated informatics challenges in home care situations (e.g. placement of devices/technologies, locating cues or reminders).

The primary goal of the videos is to provide opportunities to view and explore homes whose residents are not known to them. More than just providing experiences with home hazard assessment, these videos will be used to enable assessment of the home for such things as placement of mHealth technologies, planning for accommodations in the home that are needed to support self-care and rehearsing communication and demeanor when encountering a home that may be very cluttered or very different from the environments to which they are accustomed. Videos also provide visual cues that prepare the trainees to better participate in telehealth encounters with patients, because the trainees will better understand varied home environments.

In a pilot test, student trainees reviewed a video and case scenario; an important and unexpected barrier was their visceral response to the varying degrees of clutter and disarray. Evaluation of the simulations based on the virtual home videos will focus on several aspects including: feasibility and satisfaction with the simulation (student and course faculty), fidelity of the experience (Witmer’s Presence Questionnaire), workload (NASA-TLX) and impact.

Conclusion

This project is the first step in a modular approach to the design of many simulated home care experiences for health professional students to learn about how both health professionals and lay people use computer and information technologies in home-care and self-management. The ultimate goal is to provide easy-to-use tools to address a variety of learning needs for use by clinical experts. The pilot trial of a video supports the need to afford such guided opportunities to practice communication, priority-setting and decision-making in the home setting.

Acknowledgement: Funded by the UW-Madison School of Nursing Research Committee

Reference

Modeling Information Behavior over Time: A Colorectal Cancer Screening Case Study

Caitlin Champion, MD¹, Gonzalo G. Alvarez, MD, MPH¹, Ewan Affleck, MD², Craig Kuziemsky, BSc, PhD¹

University of Ottawa, Ottawa, ON, Canada, Northwest Territories Health and Social Services, Yellowknife, NT, Canada

Purpose

Existing modeling approaches are limited in their ability to capture the complexity of health system problems such as limited colorectal cancer screening access in a rural and remote environment¹. A combined Collaborative Information Behavior (CIB)/continuity of care framework for health care process modeling was developed and applied to identify and describe the complex health system factors contributing to screening access.

Methods

Transcripts from interviews with health care providers across the Northwest Territories, Canada (N=29) were analyzed using a combined CIB/Continuity of Care framework for qualitative content analysis²,³. Exploratory process models applying the combined framework to depict colorectal cancer screening access were developed.

Results

Within the combined modeling framework, CIB categories identified and described health care system components, while continuity of care categories were used to model care processes in screening access over time (Figure 1).

Figure 1. Process model applying combined CIB/Continuity of Care modeling framework

Information behaviors were identified as foundational to colorectal cancer screening access, and gaps in these behaviors contributing to access problems were identified and described, including siloed information systems and gaps in results reconciliation and recall, in combination with influential rural and remote contextual factors.

Conclusion

A combined CIB/Continuity of Care modeling framework can be applied to qualitative modeling of complex health system problems such as colorectal cancer screening access, identifying crucial information behaviors and system gaps where targeted informatics solutions may improve screening access and cancer outcomes. These findings may be applied to modeling of similar complex problems and inform improvements in rural and remote screening access.

References

Knowledge Management of Drug Indications – Complexities and Challenges

Christine M. Cheng, PharmD and Jeff Bubp, PharmD
First Databank Inc, South San Francisco, CA

Abstract

Health information systems use medication indication information for various applications, e.g., identifying FDA-approved and appropriate unapproved “off-label” uses for a medication, associating a drug order with an indication, indications-driven claims adjudication, patient problem list augmentation, prior authorization, and indications-based prescribing. Each use case requires consideration of many factors. We will present some of these considerations and challenges in detail and propose constructs for indications that attempt to account for these factors.

Introduction

Health information systems use medication indication information for a variety of applications. Examples of use cases for indications include identifying FDA-approved and appropriate unapproved “off-label” uses for a medication, associating a drug order or prescription with an indication, consumer labeling of prescription vials, indications-driven claims adjudication, patient problem list augmentation, prior authorization, and indications-based prescribing.\(^1,2\) Each of these use cases requires consideration of many factors, including role of therapy (e.g., treatment, prevention, diagnosis, symptom management) as well as the patient population (e.g., gender, age group, comorbidities, genomic factors) and place in therapy (e.g., first-line, second-line, adjuvant, etc.). Medication indications can also be described in multiple ways that seem divergent from one another. For example, the FDA regulatory approval (e.g., “skin and soft tissue infection” for an antibiotic) may differ from standardized terminologies (ICD-10, SNOMED CT) describing the same condition. Health care providers may use a different vernacular for use in their daily work (e.g., “cellulitis”). New terms may replace older terms (e.g., “juvenile idiopathic arthritis” instead of “juvenile rheumatoid arthritis”) in guidelines and in the literature. In addition, consumer-facing terms for indications may differ from terms used by health care providers. A knowledge base that represents the nuances and complexities of medication indication information may help support the different use cases for indications.\(^3\)

Data model requirements

We define the following model requirements: (1) ability to associate a highly granular drug indication (e.g. a patient population is specified as part of the indication) to a higher level disease concept (e.g. roll-up term), (2) ability to attribute selected drug indications with therapy intent (e.g. empiric vs definitive, preventative vs. maintenance, etc.), (3) ability to associate drug indications with patient population criteria (e.g. age, gender, disease genomics, stage of disease, etc.) when listed as part of the drug indication, (4) ability to associate the indication with standardized terminologies. Examples of drug indication modeling using different constructs will be described.

References:

Technology Usage and Preferences during Pregnancy

Sai Chennupati¹, Marian Dorst¹, Shilo Anders, PhD², Gretchen P. Jackson, MD, PhD.²
¹Vanderbilt University, Nashville, Tennessee; ²Vanderbilt University Medical Center, Nashville, Tennessee

Introduction

During pregnancy, prospective mothers and their caregivers have numerous information needs related to the upcoming birth of the child, many of which are unmet. A growing number of technologies and resources are available to support pregnant women and their caregivers. The goal of this study was to determine the technologies used by pregnant women and caregivers and their willingness to use emerging technologies to support their health.

Methods

This project was conducted as part of an ongoing longitudinal study of the information needs and information management practices. Adult pregnant women and their caregivers (e.g., spouses) seen in an advanced maternal fetal care and group prenatal care practices at Vanderbilt University Medical Center were eligible. Participants were queried about their use of text messaging, automated telephone calls, video calls, social media, and online forums, and their willingness to use these technologies to communicate with health care providers or to support their health.

Results

Twenty six individuals have been enrolled in the study. 18 pregnant women and 8 caregivers. Pregnant participants had a mean age of 32 years; caregivers, 35 years. Thirteen participants (50%) were White; 8 (31%), Black; 2 (8%), Asian; and 3 (12%), other. The mean gestational age was 24.4 weeks. The usage of and willingness to use technologies among pregnant women and caregivers are shown in Figures 1 and 2, respectively.

Conclusions

In our study, most pregnant women and at least half of caregivers used text messaging, automated calls, video calls, and social media. Over one third had used some of these technologies to communicate with their providers, but fewer had used them support their health in other ways. Over half were willing to communicate with providers and support their health through text messaging and video calls. Fewer were willing to use automated calls, social media and online forums to support their health and communicate with providers. Discrepancies between actual usage and willingness to use technologies may provide innovative ways to meet information needs during pregnancy.

Acknowledgments

This research was funded the Agency for Health Research and Quality, Grant R01 HS021496. Study data were collected and managed using REDCap electronic data capture tools hosted at VUMC.
Informatics Tools to Support Family Caregivers’ Pain Management in End-of-life Care: Current Evidence and Opportunities
Nai-Ching Chi, PhD Candidate, George Demiris, PhD, FACMI
University of Washington, Seattle, WA, USA

Background: In 2014, more than 1.6 million terminally-ill Americans received hospice care and most of them needed their family caregivers to assist with daily activities or clinical care. In 2015, an estimated 43.6 million U.S. adults served as informal caregivers for a family member. Pain management has been identified as the most challenging task by family caregivers in end-of-life care. Many argue that informatics tools such as telehealth may prove useful in assisting with pain management especially for rural and isolated caregivers. Recommendations on the design and implementation of such tools and documentation of their effectiveness are limited.

Purpose: To identify opportunities for the use of informatics tools to support family caregivers in pain management.

Methods: We conducted this study in three phases. In the first phase, we aimed to understand the magnitude of pain management challenges caregivers experience, and for that purpose we conducted a theory-driven thematic analysis of secondary qualitative data obtained from a larger clinical trial with 514 hospice caregivers. The participants were included in the analysis only if they discussed pain management difficulties with the interventionist in their interview. Two coders coded all transcripts thematically using Kelley’s “Informal hospice caregiver pain management concerns” framework. The different themes between two coders or newly identified themes against Kelley’s framework were resolved through discussion. In the second phase, we conducted a systematic review of the literature to identify existing informatics systems and applications that have been designed to target supporting family caregivers. In 2014, we searched studies from the Embase, CINHAL, Cochrane and PubMed databases using the following keywords: “telehealth”, “telemedicine”, “telecare”, “telemonitoring”, “caregiver” and “family.” We reviewed studies that described a telehealth-based intervention or application with a focus on family caregivers’ outcomes in all settings, life stages and clinical conditions. Studies were included if they used any telehealth interventions and focused on family caregivers’ outcomes. Studies were excluded if they were not published in English. In the third phase, we mapped existing interventions to identified challenges to explore existing or new ways to utilize informatics to support caregivers. In other words, we identified which of the caregiver needs had been addressed by specific interventions and if what these interventions were, as well as searched for those challenges that do not seem to have been addressed by any informatics based intervention to date.

Results: In the first phase, 15 family caregivers’ interviews were included. The prominent themes emerged from the interviews including caregivers’ physical and psychological limitation or burden, caregivers’ belief system (fear, religious belief, ethical belief or moral belief), caregivers’ inability to accurately assess pain, and ineffective communication between caregiver and healthcare provider. The results in the first phase showed that family caregivers need structured education and healthcare providers’ support in pain management. In the second phase, we conducted a systematic review of studies that employed telehealth interventions to support family caregivers. The initial search found 4205 articles. After applying inclusion and exclusion criteria, a total of 65 studies was included in this review: 52 experimental studies, 11 evaluation studies, 1 case study and 1 secondary analysis. The types of technologies utilized included video, web-based, telephone-based and remote monitoring. More than 95% of the studies found that interventions delivery through technologies significantly improved the family caregivers’ outcomes. In the third phase, mapping existing evidence to caregiver needs highlighted various areas where informatics has already played or can play a role. These include pain management education, cognitive behavioral interventions (coping, problem-solving strategies), overall support, and oversight by clinical staff (e.g. remote observation of caregiver administering pain meds).

Conclusion: Family caregivers do not receive formal clinical training but they are asked to manage patients’ pain and administer pain medications for their loved ones. Our study showed that a lack of adequate pain management knowledge as well as misperceptions about pain management prevent caregivers from managing pain effectively. Telehealth brought significant improvements to the caregivers’ outcomes, and caregivers were satisfied with the telehealth technologies. Our study results suggested that informatics technology is a promising tool to support family caregivers to be successful in pain management.
Card Sorting of Symptom Self-Management Strategies to Inform the Development of a mHealth App in underserved Persons Living with HIV

Hwayoung Cho, MSN, RN, Lena M. Milian, BSN, Rebecca Schnall, PhD, MPH, RN-BC
School of Nursing, Columbia University, New York, NY, USA

Introduction

As advances in HIV treatment have increased survival, HIV has largely become a chronic disease. At the same time, persons living with HIV (PLWH) are confronted with persistent symptoms related to HIV infection, medication side effects and comorbidities. Effective management of symptoms is essential, since it has been shown to support adherence to antiretroviral medications, improve quality of life and increase engagement with healthcare providers. In response to this need, Holzemer developed a paper-based symptom management manual with self-management strategies for 21 common HIV-related symptoms (T32NR0007081), and the manual was founded to be efficacious in a 775-person RCT over 3 months. To facilitate uptake, a web-based patient-centered tailored symptom management tool, video information provider (VIP), was developed and the feasibility of the use of the system was demonstrated in a 3 month pilot study (P30NR010677-03S1). Nonetheless, subsequent use of the symptom strategies has been very limited. The use of mobile technology has the potential to improve communication, access, and information delivery to racial/ethnic minorities. Thus, mobile health (mHealth) applications (apps) have the potential to be effective channels to disseminate the manual of symptom management in underserved PLWH. Prior to developing a mHealth app which incorporates the self-management strategies from the symptom management manual, we employed user-centered design methods to guide the information architecture of the mHealth app for symptom self-management.

Methods

We recruited 20 PLWH aged 18 and older from 1 HIV Medicaid clinic and 4 community based organizations in NYC between Dec, 2015 and May, 2016. Demographic characteristics were obtained by self-report. A reverse card sorting technique was conducted to inform the development of a mHealth app for PLWH. Users were presented with a pile of cards representing symptoms and self-management strategies. 152 self-management strategies for 13 symptoms were included in the card sorting exercise. Symptoms included: fatigue, depression, dizziness, anxiety, neuropathy, insomnia, fever, cough/breath shortness, forgetfulness, skin problem, weight loss, nausea, and diarrhea. Participants were asked to select an index card of symptoms they experienced in the past 7 days, and then place index cards of self-management strategies in priority order. Descriptive data for demographics were analyzed using SPSS. A hierarchy analysis was conducted for establishing the rank order of symptoms and self-management strategies.

Results

Our sample included 20 participants between 38-64 years of age. 12 (60%) participants were female and 8 (40%) were male. 11 (55%) participants self-identified as African American and 8 (40%) self-identified as Latino. Participants identified the following symptoms which they experienced during the past 7 days. The most frequently reported symptoms and the top 3 self-management strategies for each symptom are listed in Table 1.

Conclusion

As ensuring the needs of underserved PLWH via a participatory design process, findings from this study will be incorporated into the information architecture of a prototype of mHealth app for supporting symptom self-management. The information collected in this study will inform app navigation, hierarchical menus, and the priority order of self-management strategies.

Acknowledgements This study was supported by the Agency for Health Research and Quality (R21HS023963; PI: R. Schnall).

Reference

Why patients in online communities report not taking SSRIs

Jason H.D. Cho, M.S.1, Marina Shah, B.S. 1, Roxana Girju, Ph.D. 1
1University of Illinois at Urbana-Champaign, Urbana, IL

Abstract

Patients often post on online forums to discuss their health or their prescribed medications. We focus on why depressed patients are not taking Lexapro, in particular, why they are hesitant to take it and why they have stopped taking it. We find that while the most common reason for not taking the medication is aversion to side effects, other reasons, such as price or a previous negative experience, are also contributing factors.

Introduction

72% of patients discontinue antidepressant use within 90 days1, in spite of the literature showing that SSRIs are effective treatments. This is in spite of the literature showing these medications are effective in treating the given symptom. Motivated by this study, we leverage online communities to investigate why depressed patients are not taking medications. Researchers have utilized online communities to discover adverse drug reactions2 or to detect epidemics. In particular, we focus on one medication, Lexapro, in the context of depression treatment, and ask two research questions: First, why do patients discontinue prescribed medication? A previous study3 showed that patients stop taking psychiatric medication because 1) they experience loss by use of medication (adverse drug reactions), 2) personal meanings associated with taking medication, 3) feelings evoked by the process of taking medication and 4) perceived changes in payoff matrix (efficacy of medication). However, this was a small-scale study based on interviews. In this research we investigate whether data from online communities reflects similar patient concerns. We also investigate reasons patients are hesitant to begin taking a given medication.

Dataset and Analysis

We crawled www.healthboards.com, an active online health support community. For the purpose of this study, we looked into Lexapro, a very popular antidepressant, and took opening posts mentioning the medication from the Depression forum. There were 1,490 mentions of the drug. We sampled 300 of the mentions, and two annotators labeled forums in two ways. First, we labeled whether a forum post has mentions of ceasing to take Lexapro. Next, we looked for posts with mentions of patients' hesitancy to begin taking Lexapro. The Kappa agreements were 0.67 and 0.64 respectively. Labelers then analyzed posts on which they had disagreed, and decided on the final label.

We found 73 posts in which patients mentioned why they had stopped taking Lexapro. Of these, side effects were the single biggest reasons why they stopped taking the medication (35 posts). Ineffectiveness of medication followed, with 18 posts that mentioned this reason. Patients also indicated a desire to wean off the medication in eight of these posts. Six posts were due to changes in prescriptions, and four posts mentioned price or insurance reasons. There were other reasons, such as feeling better, or simply having no mentions on why they stopped.

There were not as many mentions (25 posts, ~10%) on why patients were hesitant to start taking medication. The predominant reasons were fear of side effects (10 posts), or previous negative experiences with Lexapro or other medications (4 posts). 8 patients mentioned searching for the best option as a reason for not taking Lexapro, and other reasons include price and uncertainty about the helpfulness of Lexapro. We see from our preliminary studies that while fear of adverse drug reactions is a big reason why patients either stop or are hesitant to take medication, perceived efficacy of the drug, and even the desire to not be on medications were also big factors. For future works, we intend on investigating an NLP technique to automatically filter out why people have stopped, or are hesitant to take medications which would allow us to scale our study to a bigger number of medications and symptoms.

References

A Single-Item Stress Scale for mHealth Interventions

Wonchan Choi, PhD¹, Sherry L. Pagoto, PhD², Edwin D. Boudreaux, PhD², Bengisu Tulu, PhD¹

¹Worcester Polytechnic Institute, Worcester, MA; ²University of Massachusetts Medical School, Worcester, MA

Introduction

Stress is associated with increased risk for obesity and cardiovascular disease¹. Mobile technologies could be an effective platform to monitor and manage stress continuously without interrupting daily routines. However, stress measurement instruments are typically lengthy and not suitable for the mobile platform. The goals of our project are (1) to develop and validate a single-item stress scale optimized for mobile delivery, and (2) to design an mHealth app interface that facilitates frequent stress data input in a timely manner with accuracy and minimal effort.

Our scale has six anchors with numerical values and visual indicators (faces that represent different levels of stress) corresponding to the six anchors. We used the item wording of a validated measure of stress symptoms, which includes a definition of stress in it: “Stress means a situation in which a person feels tense, restless, nervous or anxious or is unable to sleep at night because his/her mind is troubled all the time. Are you experiencing this kind of stress currently?”²

Methods

We recruited participants with experience in using a touchscreen-based smartphone from a university located in the northeast of the United States (n = 8, as of March 4, 2016 – study is ongoing). The study consists of two phases. In Phase I, participants answer an online survey where they match six different visual indicators (faces) we created with six anchors of stress. This task is designed to examine whether the visual indicators and participant’s perceptions of different levels of stress are aligned. Participants are then asked to list up to ten stressful events they were experiencing in their daily lives. Later, they are asked to rank the self-identified stressful events in the order of intensity of stress and rate them using the stress scale. In Phase II, we test four different mobile app user interfaces (UIs) we designed for frequent stress data input: (a) Tap-Vertical, (b) Tap-Horizontal, (c) Slide-Vertical, and (d) Slide-Horizontal. Participants are asked to input a given level of stress across different UIs as quickly and as accurately as they could using our test phones running the mobile app we developed for this experiment. After this task, they fill out a questionnaire that captures their perceptions of easiness and usability of the four UIs they tested.

Results

In our Phase I sample, a majority of the participants (7 out of 8; 87.5%) successfully matched the six faces to the corresponding levels of stress as intended. Participants identified 37 stressful events in total. A Spearman’s rank-order correlation was run to determine the relationship between the rank orders of and the ratings on the 37 stressful events. There was a statistically significant, strong negative correlation between the rank orders and ratings (r_s = - .64, n = 37, p < .001), which implied a good concurrent validity of the single-item stress scale under development.

The preliminary findings from Phase II showed that the tap-based mobile app UIs required less time for data input than the slide-based version (Tap-UIs = 2555.94 < Slide-UIs = 2747.75 milliseconds, on average). The participants rated the tap-based UIs higher than the slide-based UIs (Tap-UIs = 4.34 > Slide-UIs = 3.84 out of 5, on average) in terms of usability.

Conclusion

We will continue to collect data from a wider range of user groups to validate our stress scale. The findings from this study will provide useful insights for mHealth interventions targeting stress management in addition to other mHealth interventions involving frequent data input tasks in various contexts.

References

Crowdsourced Evaluation of Medical Texts Simplified by Medical Trainees

Yong Choi, MPH¹, Anne Turner, MD, MLIS, MPH¹, Katrin Kirchhoff, PhD¹
¹University of Washington (UW), Seattle, WA, USA

Introduction

With the drive towards greater patient engagement in healthcare, more and more medical texts (doctor’s notes, discharge summaries, etc.) are becoming available to patients. However, medical notes are complex and generally not written with patients and caregivers in mind. Medical notes are likely to include specialized medical language that is not easily interpretable by a lay person, especially those individuals with low health literacy. To address this problem, one German non-profit project has created a service for lay individuals whereby medical text is simplified into plain language by volunteer medical students [1]. Their pilot study demonstrated that the written communication skills of the students who translated medical documents on the website were enhanced. We explored whether medical trainees in the U.S. have sufficient expertise to be able to convert medical reports into understandable everyday language. The goal of this project was to determine to what extent text simplifications provided by medical trainees can improve lay persons’ understanding of medical text.

Methods

Four medical fellows and students at UW, who had completed at least one year of medical training, were recruited for the study. The medical trainee participants were given five short (5-10 sentences) excerpts from five different medical reports of various types (e.g. surgical note, procedure note, discharge note, clinic note) and were asked to express the same information in simpler, more understandable, language (e.g., 8th grade reading level). No further training was provided. We collected a total of 20 simplified texts from 5 original medical excerpts. We obtained UW IRB approval prior to initiating the study. Because the text simplifications varied greatly in their length and level of simplification, we decided to investigate which of the “translated” simplified texts were easier to understand by lay readers. To measure actual text comprehension, we created two 3-answer multiple-choice comprehension questions for each of the five original excerpts. We recruited lay participants through Amazon’s Mechanical Turk (mTurk) crowdsourcing platform. The mTurk Participants were each asked a total of 10 comprehension questions (2 questions for each of the 5 texts). The participants were randomly assigned to read either the original text or one of the simplified versions, and then asked to answer the questions. Eligibility: participants had to be 18 years or older, native English speakers, and connected to a U.S. IP address. We collected basic demographics and health literacy scores using the Brief Health Literacy Screen (BHLS) tool.

Results

179 mTurk participants completed the study. The recruited participants had a mean age (SD) of 37 (12) years old and 56% were female. The majority of the participants had at least some college education or higher (90%). The BHLS median (IQR) score was 13.0 (12.0–15.0) and the mean (SD) was 12.8 (2.1). Among the participants 167 (93.3%) had a BHLS > 9, and 12 (6.7%) had low health literacy score (BHLS ≤ 9). The percentage of correct answers was 74% for the original texts and 85.6% for the simplified texts (P=0.018). Among total 40 comprehension questions associated with the simplified texts, the correct answer rate was higher in 31 questions (77.5%) compared to the original texts. Out of the 31 questions, 20 questions showed over 15% increase in understanding but 9 questions showed less than 5% increase in understanding. Simplified versions showed lower levels of understanding by the participants in in 9 out of 40 questions (22.5%).

Discussion

Overall, the findings of this study indicate that simplifications of medical notes by medical trainees led to better understanding of medical texts. This was despite the lack of explicit task-specific training of students, and despite the fact that most mTurk participants had a fairly high literacy level. In the few cases where the simplified texts performed worse than the original texts, we found that the simplified versions either omitted critical information or contained errors. For instance, one simplified version incorrectly translated a medical jargon ‘q.d.’ on prescription (meaning one a day) to ‘twice a day’. Although improved lay person understanding can be provided through text simplification by medical trainees, training and oversight on medical simplification is recommended to ensure accuracy and consistency.

References

The Design of a Patient-Centered Personal Health Record with Patients as Co-Designers

Arlene Chung, MD, MHA, MMC,¹ ² Haiwei Chen,¹ Grace Shin, MS,¹ ² Ketan Mane, PhD,³ Hye-Chung Kum, PhD⁴

University of North Carolina at Chapel Hill School of Medicine,¹ Carolina Health Informatics Program,² Kaiser Permanente,³ Texas A&M University⁴

Introduction. The promise and potential of connected personal health records (PHRs) has not come to fruition. This may be, in part, due to the lack of user-centered design and of a patient-centric approach to curating personal health data for use by patients. Co-design with end-users could help mitigate these issues by ensuring the software meets user’s needs, and also engages patients in informatics research. Our team partnered with patients with multiple chronic conditions to co-design a patient-centric PHR. This poster will describe our experience with the co-design process, highlight functionalities desired by patients, and showcase the final prototype.

Methods. We conducted three design sessions (90 minutes per session) with patients as co-designers and employed an iterative process for software development. Patients were recruited from Chapel Hill and surrounding areas. The initial design session laid the foundation for future sessions, and began with brainstorming about what patients thought their ideal version of an engaging connected PHR would look like in terms of features and functionalities. After each software iteration, our entire design team, including our patient co-designers, was shown the prototype during a subsequent design session. Once the final prototype was developed, usability testing was conducted with patient participants. Our team then conducted a final design session to debrief about the final prototype.

Results. We started with an initial group of 12 patients (6 males) who all had diabetes and an additional comorbidity such as hypertension and hyperlipidemia. Age of participants ranged from 30-77 years with an average age of 56. The majority of participants were Caucasian with one Asian and two African Americans. Hemoglobin A1c values ranged from 6.0-9.2% with approximately half having A1c values less than the goal of 7.0%. Half the patients were aware of PHRs, majority had smartphones, and all participants had access to the Internet and used email. Two of the patients were retired engineers who had prior experience with software design. The other sessions had between 7-8 participants at each session, and 7 patients completed the 90-minute usability testing session. There was a core group of 7 patients who were engaged in the design and testing sessions throughout the entire 9-month study.

Key features of the PHR that emerged from design sessions included the following: 1) allow for annotation of data by patients (particularly important for lab values like glucose or for physical activity); 2) calendars, to do list, and reminder functions should be linked so that an entry in one of these allows for auto-population of this data within the other sections; 3) notifications whenever new data from the electronic health record or other sources are pushed to the PHR account; 4) allow for drag and drop of photos of pills/medications taken via smartphone or from other sources so that medication list has photo of actual pills or pill bottle; 5) allow for patients to customize the order of sections in the PHR dashboard so that the sections most important to the individual patient can be displayed more prominently; 6) allow for notifications from pharmacies to be pushed to the PHR (e.g. confirmation of receipt of prescription requests or alert that prescription is ready to pick up); and 7) graphical display of trends over time (patients would like to select the measures and time frames to plot for display). Patients cited the importance of data provenance so that patient-entered data vs. provider or electronic health record data could be easily differentiated. Patients also highlighted the importance of having this PHR be a “one-stop shop for all their health data” and to have meaningful data dashboards for the different types of information needed to comprehensively manage their health. Patients wished for a single PHR that could easily bring together data from multiple patient portal accounts to avoid having to manage multiple accounts and passwords. They felt that heat map displays such as those used on popular fitness tracking websites were not intuitive and that the color-coding made interpretation challenging. Participants noted that engagement in the design process made them feel that they contributed towards developing software that could not only positively impact them individually but others as well. Every patient indicated the desire to participate on future design projects. Of the 19 tasks evaluated during usability testing, only 5 tasks could not be completed (e.g. adding exercise to the calendar, opening the heat map, etc.). Patients felt that the overall PHR design was clean and aesthetically pleasing. Most patients felt that the site was “pretty easy to use” (6 out of 7). The majority of participants would like to use this PHR in the future (5) and would recommend this PHR to their friends/family to use (6).

Discussion. Involving patients directly in the design process for creating a patient-centric connected PHR was essential to sustaining engagement throughout the software life cycle and to informing the design of features and functionalities desired by patients with diabetes and other comorbidities.
Assessment of the Heterogeneity of Outcome Measurements for IT Interventions in Health Care

Tiago K. Colicchio¹, MBA, MS, Julio C. Facelli², PhD, Guilherme Del Fiol¹, MD, PhD, Debra L. Scammon², PhD, Watson A. Bowes III¹,³, MD, MS, Scott P. Narus¹,³, PhD

¹Biomedical Informatics, University of Utah, Salt Lake City, UT; ²David Eccles School of Business, University of Utah, Salt Lake City, UT; ³Medical Informatics, Intermountain Healthcare, Salt Lake City, UT

Introduction

Adoption of Electronic Health Record (EHR) systems has significantly increased in the U.S. partially due to financial incentives provided by the CMS Meaningful Use program. Despite substantial research on the impact of EHR adoption in health care, our understanding of how IT tools contribute to improving health care outcomes is still limited. A significant contributing factor to this gap is the lack of standardized measurements that could be used to compare different studies. In the present study, we analyze and categorize the measurements reported in evaluations of health IT interventions. We also identify potential ways to create a common set of measurements that could facilitate comparison among future studies.

Methods

We screened 236 studies from a previously published systematic review [1]. To create an inventory of measures that could be used in future studies we performed the following steps: (1) identified all measures used as a dependent variable and identified the targeted population; (2) conducted a bottom-up analysis grouping the measures by similarity into less specific measures (as determined by the expert opinion of the authors), creating a hierarchy that excluded unique measures that were used in only one study or could not be grouped into other more general measure; and (3) grouped the measures from the highest level of the hierarchy into categories that represent the types of measurements.

Results

From Step 1 above we identified 429 unique measures and from Step 2 we excluded 239 (56%) unique measures, leading to the inclusion of 190 measures at the most specific level of our hierarchy. The 190 measures were grouped into 66 less specific measures in the highest level of the hierarchy, and then further grouped into 15 types of measurements. The most common types of measurements were “test or procedure ordered as preventive care”; followed by “optimal care documented in the patient’s EHR” and “appropriate use of pharmacotherapy”. Figure 1 provides an example of the hierarchy for one measure: “appropriate use of antibiotics”.

Discussion

Our analysis demonstrates that the measures used in health IT adoption research vary substantially, creating obstacles to the comparison of health IT outcomes. Our hierarchy was used to create a taxonomy of commonly used measures that may help researchers to identify measures and gaps in their measurement approaches. The taxonomy will be published in a separate study; its first version can be accessed at https://goo.gl/EiwyKR. The use of a previous systematic review is a limitation to the identification of measures. We expect to address this limitation by eliciting more measures from informatics experts and report our results in a future study.

Acknowledgement

This project was supported by Intermountain Healthcare, Salt Lake City, UT. JCF has been partially funded by National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number 1ULTR001067.

References

A Qualitative Analysis of Electronic Clinical Quality Measures Development and Data Validation

Nicholas V. Colin, MA, Raja A. Cholan, BS, Shelby J. Martin, MS, RD, Bhavaya Sachdeva, MPH, David A. Dorr MD, MS
Oregon Health and Science University, Portland, OR, United States

Introduction. Electronic Clinical Quality Measures (eCQMs) are tools to measure and record the quality of care being provided by healthcare professionals. The precise implementation of eCQMs can identify care gaps and produce accurate reports for incentive programs such as Meaningful Use or the Physician Quality Reporting System. However, it can be difficult to determine whether eCQMs are implemented accurately, as storage, retrieval, and processing of key data varies substantially by practice and provider. We utilized a six step process for developing and validating eCQM reports (1. identifying needs/assessing requirements for the measure, 2. initial code/concept mapping, 3. team code review, 4. implementation of non-validated measure in test environment, 5. validation and revision, and 6. implementation as validated measure). This allowed us to improve our implementation and increase trust of the data. Our objective is to describe our approach to help improve accurate eCQM implementation for others trying to do this work.

Methods. Our group used prior work to encode the six step process during the development and implementation of three eCQMs (Aspirin use when appropriate, CMS164v4; Controlling High Blood Pressure, CMS165v4; and, Cholesterol management, CMS347v0) as outlined in Table 1 below. First, we identified the need and assessed requirements for each measure by reviewing the official measure specifications from the National Quality Forum, and compared the criteria to any previous versions of the measure. Next, we coded the necessary logic and mapped aggregate concepts, reviewed the code, implemented the measure in a test environment, and validated the measure by conducting chart review of patients’ records who were returned in the denominator/numerator.

Results. Table 1 highlights the process and changes by measure. Review of the aspirin code identified an error in the numerator coding in which medications were mislabeled as antithrombotic. Validation and revision of the patient charts from the blood pressure denominator revealed an error in the lookback period for most recent blood pressure recording. Review of the cholesterol measure code identified missing criteria making the denominator too broad; this correction lowered the count of patients in the denominator from 3,903 to 1,565. No coding changes were recorded for step 1 (identifying needs/assessing requirements for the measure) or step 6 (implementation as validated measure) for any of the three eCQMs (Aspirin, Blood Pressure, and Cholesterol).

Table 1. Six steps of eCQM development and number of coding changes documented

<table>
<thead>
<tr>
<th>Steps of eCQM Development</th>
<th>Aspirin (# of coding changes)</th>
<th>Blood Pressure (# of coding changes)</th>
<th>Cholesterol Measure (# of coding changes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2: Initial code/concept mapping</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Step 3: Team code review</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Step 4: Implementation of non-validated measure in test environment</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Step 5: Validation and revision</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion. We codified the six step process for implementing and validating eCQMs; we found following and documenting this process improved implementation. The validation process provides an opportunity to identify areas for improvement including coding errors, gaps in care that exist in current workflows, and value set discrepancies used in measure definitions for eCQMs. Through this process, we were able to identify errors affecting the accuracy of our eCQM implementation. We were able to improve the overall performance and accuracy of the eCQMs, and once corrections were implemented we were able to validate against patient-level data to improve our trust and confidence with the data.
Usage of Electronic Online Care Process Models for Learning at the Point of Care

David A. Cook, MD, MHPE1,2,3 Laurie J. Pencille, CCRP1, Kristi J. Sorensen, MS1, Deborah J. Rhodes, MD1,2

1Knowledge Delivery Center, Mayo Clinic College of Medicine, Rochester, MN; 2Department of Medicine, Mayo Clinic, Rochester, MN; 3Mayo Clinic Online Learning, Mayo Clinic College of Medicine, Rochester, MN

Introduction

Clinicians constantly seek and apply new information in response to patient-focused questions.1 Electronic clinical decision support systems and knowledge resources help to efficiently answer these questions.2 AskMayoExpert is a novel multifaceted knowledge resource, integrated with the electronic medical record, providing concise information on >1700 clinical topics and >140 care process models (CPMs).3 CPMs reflect institutionally-approved evidence-based standards for managing common clinical problems. These are implemented as flowchart-like interactive algorithms highlighting key decision points. At each decision point, the CPM provides relevant information about the problem (e.g., disease classifications and risk factors), suggests additional information to be obtained (e.g., interview questions, exam elements, or laboratory tests), and recommends next actions. All CPMs are developed and approved by a committee of topic experts who employ state-of-the-art evidence and clinical guidelines. Generalist clinicians may find CPMs particularly useful.

We sought to determine how often clinicians (physicians, residents, nurse practitioners, and physician assistants) access CPMs; the relative usage by generalists vs non-generalists; and how often clinicians return to the same topic.

Methods

We determined usage from January 2014 – June 2015 using computer logs. We compared usage rates for generalists vs non-generalists using the Wilcoxon rank-sum test, and enumerated the number of recurrent views of the same topic by a single user.

Results

We have developed 142 CPMs. Over the 18-month period, 4076 clinicians (61% of approximately 6700 clinicians with access) viewed CPMs a total of 81,853 times. Usage varied from 1 to 809 views per person, and from 9 to 4621 views per CPM. We were able to determine the clinical specialty for 2754 clinicians. Among these, generalists (N=1415) used CPMs much more often (median 7, mean 31.4 views) than non-generalists (N=1339; median 2, mean 6.9; p<.0001). 2116 clinicians (52%) returned to the same CPM topic more than twice, with an average of 7.8 views per topic (maximum 195). 733 clinicians revisited the same topic more than 5 times, and 168 revisited more than 20 times. Generalists revisited topics more often than non-generalists (mean 8.8 vs 5.1 views; p<.0001).

Conclusion

Electronic care process models are highly used, particularly by generalists. A subset of clinicians uses them with very high frequency. Clinicians commonly revisit the same topic. Whether revisiting a topic is desirable (i.e., reflecting high perceived utility) or undesirable (i.e., reflecting failure to retain information) merits further study. We are now developing and evaluating additional specialist-focused content.

References

An Automatic Data Mining Approach Using Linked Data Technologies
Brett F. Cropp MS\textsuperscript{1}, John B. Coles PhD\textsuperscript{1}, Karen M. Meess\textsuperscript{1}
\textsuperscript{1} CUBRC Inc., Data Science and Information Fusion Group, Buffalo, NY

The traditional relational database structure loses value when working with high volume, multi-dimensional data derived from multiple sources. Misalignment of data is exacerbated due to proprietary structures that do not adhere to common data models or standards, and through several iterations of schemas, typical in the healthcare industry. One of the most significant challenges with large disparate datasets common to healthcare research and operations is that valuable information is difficult to identify, contextualize, and analyze. Techniques in the healthcare domain using broad technical data mining or the Knowledge Discovery in Databases (KDD) paradigm have been shown to provide immense potential and value\textsuperscript{1}. One particular approach for iteratively mining data, called Sample, Explore, Manipulate, Model, and Assess (SEMMA), has been used to investigate medical conditions, including healthcare-acquired infections\textsuperscript{2}. However, this approach is difficult to implement for large, dissimilar datasets\textsuperscript{2} and thus requires solutions that are both data source agnostic and scalable.

We implemented a multi-stage, automatic data mining approach that leveraged Linked Data technologies. Treating data as an interconnected web by using an ontological framework provides the ability to search information across sources while maintaining key details about an entity (e.g., individual, group, or agent of interest). Our approach used Resource Description Framework (RDF) and built upon a carefully designed and extensible mid-level ontology to align and semantically enhance data across multiple domains. In RDF, all information is represented as triples (\textit{Subject, Predicate, Object}), each of which represents a statement or assertion about something. Such NoSQL databases are flexible by having a single data representation, rather than constraining data to table models. This also allows for flexible data storage methods, as highlighted by integration with open single-instance (Apache Jena) and cluster-based (Rya with Accumulo) triple stores. By removing dependencies on traditional table models and specific sources, and applying an extensible programmatic workflow, we built an analytic framework that was shown to be generalizable across multiple domains and use-cases, with an initial focus on healthcare applications.

Our mechanism identified key information related to a base RDF graph pattern constructed in SPARQL. This pattern may represent a patient condition, particular event, relationship, or any combination thereof (e.g., patients having a diagnostic code for diabetes mellitus type II). Subsequently, the network surrounding this base graph was sampled using near-graph searches, with connected resources (e.g., John Doe) and literals (e.g., John Doe’s HbA1c test result) undergoing automatic data-type identification, classification, and contextualization. This information-agnostic near-graph search and sampling technique was then applied to different cohorts of individuals, who were compared using a predefined set of standard descriptive and inferential statistical procedures for pairwise or group comparison. Each statistical comparison was weighed by a significance factor and the cumulative discrepancy between groups was calculated using Bayesian inference, isolating a similarity metric for each information object. Finally, a threshold or confidence level was set for automatic feature selection into multivariate models including logistic regression, clustering (\textit{k}-means) and supervised learning techniques (support vector machines).

This approach was applied to several data sources, including clinical and claims. We ingested protein biomarker (nuclear translocation) data derived from allogeneic hematopoietic cell transplantation (allo-HCT) patients to identify correlations with the acquisition of acute graft-versus-host disease (GvHD). We also partnered with an insurance company to parse several million claims records, isolating correlates of case management outcomes among patients with high-risk diseases and informing patient-specific interventions and risk modeling.

Future work includes implementing time-series based resolution to identify common patterns that occur in sequence or within a particular time period. Medical data sets can be enhanced by linking to outside sources, such as social media and purchasing activity, to gain a more complete picture of individual behavior and environmental factors. A limitation with a naive approach to data mining is that certain information requires further contextualization and classifiers should be developed to provide higher-level computed information, where appropriate.

References

Developing Analytics for Wheelchair Sensors to Improve User Health

Rui Song\(^1\), Dorothy W. Curtis, M.Sc\(^1\)
\(^1\)Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

Abstract

Why can’t wheelchairs be more intelligent? This was the rationale behind integrating a contact-free sensor system with a wheelchair and subsequent deployment at The Boston Home. We are analyzing the resulting data for both long-term trends and real-time alerts. Applications include prevention of heat stress and pressure sores, and detecting sleep apnea by analyzing information from environmental sensors, accelerometers, and respiration waveforms. This has the potential to improve user health and autonomy.

Introduction

Common issues faced by long-term wheelchair users include inactivity-induced pressure ulcers, irritation from dermal sensors, and undetected sleep apnea; a comfortable and intelligent wheelchair could solve these issues. Dorothy Curtis of MIT CSAIL’s Assistive Devices for Healthcare Group and Diego Arias, Professor Esteban Pino and Professor Pablo Aqueveque from Universidad de Concepcion built a contact-free sensor system for a wheelchair to collect environmental information, motion data, and vital signs [1]. The system was deployed at The Boston Home, a care facility for patients with multiple sclerosis and other neurodegenerative diseases. The first phase of this project evaluated whether valid data could be collected while minimizing patient discomfort. Now we are developing algorithms and visualization methods for the collected data to create a retrospective summary of trends in user activity and health, and a real-time alert system.

Technical Approach

The first objective was to compile a user’s historical data into a concise summary of the following parameters: outdoor activity, general motion, pressure ulcer prevention activities, and heart rate abnormalities. The wavelet method was used to separate the ballistocardiogram signals into respiratory rate and heart activity, and remove noise generated from patient movement [2]. For each week of data, statistical baselines were established for every parameter. The summary then displays these baselines for the user-selected week along with the daily value of these parameters. The summary can also display trends in parameters over a period of several weeks. The primary user of this summary visualization is the caregiver, who may leverage this information to recommend treatments. The second objective was to collect vital-sign and environmental information to create a real-time alert system. The first application of the alert system is to alert the user to possible heat stress situations and remind the user to tilt to avoid pressure ulcers. Analysis of the respiratory and cardiac signals can give the caregiver and the user information about a potential sleep apnea episode, and, possibly other heart rate abnormalities, through the use of machine-learning classification algorithms [3].

Conclusion

The overall objective is to improve upon an existing contact-free wheelchair sensor system by providing powerful vital signs, environmental and motion data analysis and visualization. In the next phase we will evaluate the effectiveness of this method in improving wheelchair user health and autonomy by enabling them to avoid heat stress exhaustion, reduce the incidence of pressure sores, and be alerted to potential sleep apnea episodes.

References

Integrating Electronic Death Registration with Electronic Health Records Systems: Results of a Trial Use of the VRDR Profile

Vira Danak MS¹, Glenna Gobar DVM MS¹, Michelle Williamson², Michael Hogarth, MD¹
¹UC Davis Health System, ²Center for Disease Control and Prevention

Abstract
The collection of death related information is a long-standing process in society. Vital statistics derived from this process have been critical in characterizing jurisdictional mortality for over 300 years. The evolution toward electronic vital record systems and the increasing adoption of electronic health record systems have created an opportunity to dramatically decrease the latency of vital event reporting through integration of EHRs and death registration systems. This research studied the feasibility of using the Integrating the Healthcare Enterprise (IHE) Vital Records Death Reporting (VRDR) profile in successfully enabling EHR-based capture of medical certifier sourced data into an electronic death registration system (EDRS) as well as jurisdictional reporting of death events to the National Center for Health Statistics (NCHS).

Introduction
We report early results of a pilot in which the IHE VRDR profile was used to integrate the California Electronic Death Registration System (CA-EDRS) with a broadly used enterprise level EHR (Epic Corp, Verona WI) to facilitate the capture of medical information for death certificates in CA-EDRS, as well as the submission of death certificate information to NCHS.

Methods
We developed a module in our EDRS system that provides a secure data collection web-window that is provided to an EHR user within the EHR client. The VRDR profile involves an EHR user triggering a request to an external system using the IHE Remote Form for Data Capture (RFD) profile. This request includes an attached death information clinical document architecture (CDA) XML file. Our EDRS system responds by creating a new record with the received CDA file information, and responding with a private record-specific URL that is rendered by the EHR client, thus providing a data entry form to the medical certifier who is using the EHR. Subsequent to the successful completion of the death certificate and registration by the state, a process is triggered with CA-EDRS that constructs an HL-7 v2.5.1 message using the VRDR profile specification and immediately submits this message to the NCHS using an on-premises PHIN MS node.

Results
We found that the VRDR profile, as currently specified, was able to support the medical certifier data capture workflow as well as the jurisdictional reporting to NCHS. This was satisfactorily validated by the IHE at the 2016 IHE Connectathon (Cleveland,OH). We found that an EHR user with limited experience could complete the medical portion of the death certificate in as little as 10 minutes.

Conclusion
The IHE VRDR profile is a key specification in the integration of EHRs and EDRS and can reduce the latency of mortality reporting. Improving the exchange of death information using national standards will strengthen the vital statistics infrastructure for public health reporting and public health surveillance.

References

Funding Acknowledgement: This work is funded by CDC and NCHS through PCOR Trust Funding for Improving Mortality Data
Analysis of Heparin and Aspirin Administration on Acute Ischemic Stroke from International Stroke Trial (IST)
Yifang Dang, Sahitha Gajjala, M Pharm, Craig W. Johnson, PhD
The University of Texas Health Science Center at Houston (UTHealth), School of Biomedical Informatics, Houston, TX.

Introduction
The original study done on the IST dataset\(^1\) reported that early administration of aspirin is not effective in decreasing death in acute ischemic stroke (AIS). Our specific aim was to test, via secondary analysis of the dataset, the research hypothesis that; by statistically controlling and testing for age groups (below and above 70), gender, and countries; aspirin, as well as heparin, would be found to be effective on recurrent AIS or death rate within 14 days.

Analysis Procedure

![Hierarchical Binary Logistic Regression Diagram]

Results
The total IST database sample contained 19,435 patients. Binary hierarchical logistic regression analyses converged for 19,117 patients in the analysis on recurrence of ischemic stroke within 14 days and 19,340 patients in the analysis on death within 14 days. Odds of recurrence of ischemic stroke within 14 days were multiplied by 0.541, 95\%CI = (0.439, 0.667) for aspirin and 0.74, 95\%CI = (0.605, 0.905) for heparin administration, controlling for age, gender, and countries. Odds of death within 14 days were multiplied by 0.861, 95\%CI = (0.784-0.946) for aspirin and 0.983, 95\%CI = (0.895-1.080) for heparin administration and by 2.20, 95\% CI = (1.97, 2.45) for age above 70, controlling for age, gender, and countries. Aspirin by heparin interactions were not statistically significant. Countries were significantly associated with both AIS recurrence and death but gender was associated with neither recurrence nor death.

Conclusion
In contrast to results of the original study showing aspirin administration not significantly associated with reduced deaths within 14 days, we observed significant association of aspirin administration with both AIS recurrence and death within 14 days, while statistically controlling gender, countries; and testing and controlling for age above/below-70. Furthermore, age above 70 was a significant risk factor for death within 14 days. For heparin administration, which had significantly reduced AIS recurrence at doses 5000 IU and 12,500 IU, but not likelihood of death within 14 days, our results were similar to those of the original study.

References
Assessing Healthcare Workforce Training Needs for Health Information Technology-Enabled Healthcare Transformation

Raven David, MPH1; Morgan Moy, MPH1; Virginia Lorenzi, MS, CPHIMS, FHL71,2; Andrew Flatgard1; Angie Lee1; Bruce Forman, MD1,2; Lucy Appert, PhD1; Jennifer Ringler, BS1; Gilad J. Kuperman, MD, PhD1,2; Victoria Tiase, MSN, RN2; Rita Kukafka, DrPH, MA, FACMI1; 1Columbia University, New York, New York; 2New York-Presbyterian Hospital, New York, New York

Problem

Our nation’s healthcare system is changing at a rapid pace accelerated by the American Recovery and Reinvestment Act (ARRA) and the Affordable Care Act (ACA). Transformative health care delivery programs depend heavily on health information technology (HIT) to improve and coordinate care, maintain patient registries, support patient engagement, develop and sustain data infrastructure necessary for multi-payer value-based payment, and enable analytical capacities to inform decision making and streamline reporting. Additionally, EHR technology is now available to over half of the nation’s eligible providers participating in the Meaningful Use program. The accelerated pace of change from new and expanding technology will continue to be a challenge for preparing a skilled workforce. Therefore, we seek to investigate what does the healthcare workforce need to know to enable healthcare transformation as health information technology and healthcare delivery systems evolve.

Purpose

As part of the U.S. Department of Health and Human Services’ efforts to create an interoperable learning health system that achieves better care, smarter spending, and healthier people, the Office of the National Coordinator for Health Information Technology (ONC) funds programs in workforce training. As one of the awardees for this two-year program, Columbia University has performed a training needs assessment to inform the design of a training curriculum for incumbent healthcare workers for the rapidly changing landscape of healthcare and HIT.

Methods

We conducted 14 interviews with 16 key informants that are senior leaders in their respective organizations and serve on our advisory board. These informants represent a broad expertise across population health, care coordination and interoperable health IT systems, value-based care, healthcare data analytics, and patient-centered care. These interviews were transcribed and a qualitative coding and analysis was performed. A quantitative survey to complement qualitative findings is being administered to managers and their workforce to obtain ground-level data to inform tailored content and delivery of training. This protocol was approved by Columbia University IRB.

Results

Preliminary results from the qualitative surveys indicate that the trainings should contextualize why and how healthcare workers’ actions and responsibilities have macro-level effects. For example, it is important for a physician to understand how their interactions with individual patients relate to population health and payment initiatives. Respondents also expressed a need for training that clearly conveys how an individual worker’s job functions and responsibilities are interconnected with other healthcare jobs. With a paradigm shift towards team-based care, it is important for a nurse to understand how a community health worker can contribute valuable patient information, for example. Respondents also expressed that with this paradigm shift and the blurring of roles, training should be tailored to job function rather than credential (e.g., nurse, physician, etc.). Insight was provided on which roles should be targeted for training, what should be trained, and how training can be disseminated.

Conclusion

There is an enthusiastic demand from senior leadership for highly targeted HIT training for the healthcare workforce in this new era of healthcare. The quantitative survey currently being administered will further identify specific areas of training for various types of healthcare workers. With a dynamic landscape of healthcare, a challenge is to forecast what future training needs will be and balance them with the current needs to train the workforce. Additional analysis of the qualitative and quantitative surveys is being conducted and will be reported.
Automatic Extraction of Maximum and Recommended Drug Dosage Information from DailyMed Database

Lalindra De Silva, MS\textsuperscript{a,c}, Olga V. Patterson, PhD\textsuperscript{a,b}, Scott L. DuVall, PhD\textsuperscript{a,b}

\textsuperscript{a}VA Salt Lake City Health Care System; \textsuperscript{b}University of Utah, Salt Lake City, UT

Abstract
Medication dose depends on a range of factors, including drug class, drug form, patient age and weight, and frequency of administration. Detailed information on different dose requirements for each drug is described in the drug label. Due to its complexity and variability across a large number of drugs and manufacturers, this information is stored as unstructured text in public and proprietary sources. We developed a two-layered approach to extract recommended and maximum dosage information for a large number of drugs by: 1) retrieving drug information from a public database, and 2) applying a rule-based Natural Language Processing (NLP) pipeline to perform extraction.

Background
Medication data quality issues can have a lasting detrimental impact on accuracy of computerized decision support and potential research studies that utilize prescription records in modeling patients' state of health. The incorrect dose of the prescribed or administered drug is one of the potential types of errors. A simple but very effective way of detecting erroneous entries in large databases is to identify outliers using recommended and maximum doses. Detecting the individual maximum and the usual recommended dosages for a given drug is challenging because it depends on whether the drug is administered to a child or an adult, the form of the drug (i.e., tablet, capsule vs. liquid), the dosage of an individual unit (200mg tablet vs. 400mg tablet), etc. Using absolute doses that disregard variability across individuals is a potentially effective alternative to individual doses in identifying outliers. Previous studies\textsuperscript{1} have explored related approach using proprietary full-text drug monographs. We present a methodology where a publicly available database is used to extract dosage information for a large number of drugs.

Methods
We utilized the NIH DailyMed drug database (http://dailymed.nlm.nih.gov/dailymed/index.cfm) through RESTful web service to collect 50,028 drug names using the RxNorm mapping files. For each drug, we queried the web service and retrieved an HTML response. The response was parsed using the Jsoup Java library and the free text under the “Dosage and Administration” section heading was extracted. This resulted in free text entries for 47,088 drug names, leaving out the ones, which did not contain the aforementioned section heading.

We developed a dose extraction NLP pipeline using Leo architecture based on Apache UIMA framework that takes in the free text entry for each drug and identifies mentions of several dosage annotation types.\textsuperscript{2,3} The NLP system annotates free text using three dosage annotations: 1) SimpleMention, 2) RecommendedDosage and 3) MaximumDosage. The SimpleMention type identifies phrases of dosage mentions using regular expression patterns anchored around terms such as “capsules, tablets, doses”, quantifier terms such as “mg” and temporal terms such as “daily, weekly” (e.g., ‘100mg a day’). The remaining two annotation types augment these mentions, looking for modifier phrases around the context of SimpleMentions such as “maximum”, “max”, “do not take more than” for maximum dosages, and “recommended”, “usual” for recommended dosages. The system detected 139,453 mentions, of which 15,769 were MaximumDosages, 12,629 RecommendedDosages, and 111,055 SimpleMentions.

Discussion
The dose information resulted from this work can be used for as the initial filtering criteria to detect erroneous medication entries in electronic medical records. Additional development is needed to extract dosages at a more granular level, relating them to patient characteristics (age, weight) and medication variables (form, administration frequency). Since most publicly available drug information repositories are geared toward individuals (clinicians or consumers), the developed system bridges the gap and makes dosage information available for data quality auditing.

Acknowledgements
This work was performed using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204.

References
Effects of Meaningful Use of EHRs on ED Clinical Workflow

Courtney Denton, BA1, Gloria Nimo, MSN2, Jason Shapiro, MD2, Thomas G. Kannampallil, BA3, & Vimla L. Patel, PhD1
1The New York Academy of Medicine, New York, NY; 2Mount Sinai Medical Center, New York, NY; 3University of Illinois at Chicago, Chicago, IL

Introduction
While electronic health records (EHRs) have had positive impacts on clinical practice, research has also highlighted challenges regarding their use including usability issues1, increased operation and maintenance costs2, and limited integration and support for clinical workflow activities3. To encourage widespread adoption, financial incentives have been offered to healthcare providers for the “meaningful use” (MU) of certified EHRs as they document and manage patient care information. We report on the impact EHR usage and MU criteria compliance has on clinical workflow in the emergency department (ED).

Methods
Participant observations and semi-structured interviews were conducted at Mt Sinai Medical Center and NewYork-Presbyterian Hospital EDs. During two 10-day periods, 16 physicians and 12 nurses were shadowed during their clinical shifts to capture clinician interactions and clinical activities involving EHR use using field notes. Twenty semi-structured interviews with ED clinicians were conducted to gather perceptions of EHR use and the impact of MU criteria compliance on efficiency, quality and safety of patient care. Clinical workflow-based MU considerations for EDs included door to diagnostic evaluation time and admit decision time. Clinicians were specifically asked how these MU criteria supported (or did not support) their clinical work. Interviews were transcribed, and, in collaboration with clinician collaborators, responses were coded by two researchers using Nvivo software. All coding disagreements were discussed and resolved to generate a more standardized coding scheme.

Results and Discussion
Shadowing data highlighted the sequence of steps, and the interwoven nature of EHR use during clinical workflow at two sites, where different EHRs are used. Three main themes emerged during analysis of interviews: (a) nature of EHR use, and its primary role as an information seeking and documentation tool in ED practice, (b) the influence of EHR usability on the inefficiencies in the workflow, and (c) lack of clinician awareness surrounding MU criteria compliance.
While some clinicians were aware of potential effects of MU in increasing efficiency, others merely referred to the positive aspects on EHR use and workflow, where access to a more comprehensive patient history in improving accuracy of care and tracking the status of patient care were considered important. At both academic hospitals, more physicians identified the positive rather than problematic aspects associated with EHR and MU. Clinicians’ reliance on the use of heuristics to circumvent problems in dealing with EHRs has implications for potential threats to patient safety and to the quality of care.

Conclusions
Compliance to MU criteria impacts clinical workflow. In two academic environment studied, clinicians perceive the impact to be more positive with less awareness of negative effects across two different platforms. Literature reports the opposite findings from small primary care practices. To what extent the perceived advantages correlate with the real work in the ED is under further investigation.

Acknowledgements
This project was supported by grant number R01HS022670 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors, and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

References
Methods to Measure and Improve the Quality of Large Scale Health Data: An Application in Public Health Surveillance

Brian E. Dixon, MPA, PhD1,2,3, Jon Duke, MD, MS,2 Shaun J. Grannis, MD, MS2,4
1Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; 2Regenstrief Institute Center for Biomedical Informatics, Indianapolis, IN; 3Center for Health Information and Communication, Health Services Research & Development Service, Veterans Health Administration, U.S. Department of Veterans Affairs, Indianapolis, IN; 4Indiana University School of Medicine, Indianapolis, IN

Introduction

Nearly all of the myriad activities (or use cases) in clinical and public health (e.g., patient care, surveillance, community health assessment, policy) involve generating, collecting, storing, analyzing, or sharing data about individual patients or populations. Effective clinical and public health practice in the twenty-first century requires access to data from an increasing array of information systems, including but not limited to electronic health records. However, the quality of data in electronic health record systems can be poor or “unfit for use.” Therefore measuring and monitoring data quality is an essential activity for clinical and public health professionals as well as researchers.

Methods

Using the Health Data Stewardship Framework [1], we will extend Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES), a software package published open-source by the Observational Health Data Sciences and Informatics collaborative (OHDSI; www.ohdsi.org) to measure the quality of data electronically reported from disparate information systems. Our extensions will focus on analysis of data reported electronically to public health agencies for disease surveillance. Next we will apply the ACHILLES extensions to explore the quality of data captured from multiple real-world health systems, hospitals, laboratories, and clinics. We will further demonstrate the extended software to public health professionals, gathering feedback on the ability of the methods and software tool to support public health agencies’ efforts to routinely monitor the quality of data received for surveillance of disease prevalence and burden.

Results

To date we have mapped key surveillance data fields into the OHDSI common data model, and we have transformed 111 million syndromic surveillance message segments pertaining to 16.4 million emergency department encounters representing 6 million patients for importation into ACHILLES. Using these data, we are exploring the existing 167 metrics across 16 categories available within ACHILLES, including a person (e.g., number of unique persons); and observation period (e.g., Distribution of age at first observation period). Syndromic surveillance (SS), however, is driven largely by monitoring patient stated chief complaints (non-standard free text clinical data) in addition to coded diagnoses. Consequently, ACHILLES must be extended to maximally support use in analyzing SS datasets.

Conclusion

This work remains a work-in-progress. Over the coming year, we will not only explore existing ACHILLES constructs using real-world public health data but also introduce new functionality to explore 1) patient demographics; 2) facility and location (e.g., emergency department where care was delivered); and 3) clinical observations (e.g., chief complaint). The design and methods for examining these aspects of surveillance data will be included on the poster, and they will be made freely available for distribution with a future instance of the ACHILLES software. We ultimately envision these tools being available for use on platforms such as the CDC’s Biosense – open to all local and state health agencies as a one-stop portal for surveillance data analysis – or research environments where they can be used to examine and improve the quality of data output from informatics systems.

References

Proposal of Automatic Patient Authored Text Analysis to Measure Patient Engagement

Gabriel N. Ribeiro, BSc¹, Marcia Ito, PhD²
¹IBM Software Laboratory, São Paulo, SP, Brazil; ²IBM Research, São Paulo, SP, Brazil

Introduction

Chronic diseases cause more deaths than all other causes combined, and are projected to increase from 38 million in 2012 to 52 million by 2030¹. Treatment requires behavior change, and patient engagement in chronic diseases continues to be one of the main concerns in the healthcare field worldwide². Evidences suggests that patients who are engaged tend to use fewer health care resources, make better decisions and have better health outcomes³. Measure non-engagement behavior and analyze its causes are important to identify patient in a group-risk and to develop specific interventions²³.

Assessment of Patient Engagement

Patient engagement is explained by models of behavior change. These models have in common the patient's perceptions, such what they think about their doctors, disease, treatments or themselves². First step of our method is to choose one appropriated model of behavior change. Then, data from one patient is collected from a variety of data sources (e.g: social medias, online health communities, etc) and classified according to the patient's perceptions and polarity (as positive or negative to engagement), using pre-trained machine learning models. The system uses the classified data to calculate the engagement degree for the patient. The calculation is the difference of all positive and negatives classified perceptions, with an optional weight factor for each perception for adjustment purposes, according to its relevance for engagement. Negative perceptions of this patient are associated with all available strategies to address non-engagement behavior, such as health literacy or self-efficacy intervention. The winner strategy is the one with more negative perceptions. Patient data is also used in this calculation to adjust each case, considering age, sex or any other relevant variable.

In the final step, the system presents a quantitative engagement degree with the top-N negative perceptions and possible strategies that can maximize the intervention outcomes for the patient. All the classified data, engagement degree and calculated strategies are stored on engagement database, linked to the patient on patient database.

Conclusion

Our method proposed a solution to assess patient engagement and present a calculated intervention strategy. This method can speed up engagement diagnosis because it's independent of medical consultation. Also, it does not depend on an individual and biased doctor evaluation, can be cheaper than sensors or mobile solutions Future research focuses on clinical trials to compare this method efficiency against traditional engagement assessment methods.

References


1375
GWAS Finder: Search Engine for GWAS Datasets in Biomedical Literature

Xiao Dong, M.D., M.S.1, Yaoyun Zhang, Ph.D.1, Hua Xu, Ph.D.1
1School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction

Building a new technical infrastructure for researchers to efficiently find and cite datasets is important. The aim of this study was to develop a scalable approach that can automatically scan millions of scientific publications and identify dataset mentions. In order to fulfill this final goal, we started our study by searching literature using attributes representing the underlying datasets. Genome-Wide Association Studies (GWAS) was chosen as a case study. A corpus of biomedical literature from GWAS catalog was annotated with essential attributes that are most representative of GWAS datasets, on the basis of which an automatic attribute identifier was built. Identified GWAS literature was then indexed with their dataset attributes, in order to establish a search engine, GWAS Finder, for GWAS datasets in literature.

Method

300 GWAS abstracts randomly selected from GWAS Catalog were annotated with eight essential attributes of datasets, including trait, platform, source, ethnicity, stage, sample size, case size and control size. A hybrid named entity recognition model was generated by combining conditional random fields and heuristic rules. Then a total of 13,196 GWAS abstracts collected from PubMed and GWAS Catalog were indexed using their identified dataset attributes, upon which the GWAS Finder was built for searching GWAS datasets in literature.

Results

We evaluated the performance of GWAS Finder by comparing the search results from GWAS Finder to those from GWAS Catalog and PubMed. Each performance was the average score of ten randomly selected queries. The comparisons between GWAS Finder and PubMed are shown below in Table 1, and the comparisons between GWAS Finder and GWAS Catalog are illustrated in Table 2. PubMed does not provide a search function for case size so we didn’t evaluate the performance of PubMed using “Trait + platform + case size” combination as the query type.

Table 1. GWAS Finder VS. PubMed (%):

<table>
<thead>
<tr>
<th>Query</th>
<th>GWAS Finder</th>
<th>PubMed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Trait</td>
<td>83.29</td>
<td>75.80</td>
</tr>
<tr>
<td>Trait + platform</td>
<td>83.07</td>
<td>89.42</td>
</tr>
<tr>
<td>Trait + platform + case size</td>
<td>92.71</td>
<td>75.00</td>
</tr>
</tbody>
</table>

Table 2. GWAS Finder VS. GWAS Catalog (%):

<table>
<thead>
<tr>
<th>Query</th>
<th>GWAS Finder</th>
<th>GWAS Catalog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Trait</td>
<td>93.40</td>
<td>89.00</td>
</tr>
</tbody>
</table>

Discussion

There are different ways for users to search for biomedical datasets. Searching literature for biomedical datasets, which this study aims to address, is an important aspect of data discovery process. Our future work would include efforts towards another direction, searching datasets from data repositories using similar strategies with which we used in this study and building final linkages between literature and data repositories. This switch indicates the broad uses of text mining methods in building data discovery index.

Conclusion

Our study demonstrates the potential application of text mining methods in building the data discovery index. It can create a better index of literature linked with their underlying data sets, thus improving data discoverability.
Utilizing a State-level, Integrated, Discharge Database to Monitor the Trend in Potentially Preventable Hospitalizations for Congestive Heart Failure

Riddhi P. Doshi, MBBS, MPH1, Gregory Matthews, PhD2, Garth Graham, MD,MPH3, Robert H. Aseltine Jr., PhD1

1University of Connecticut Health Center, Farmington, Connecticut, USA; 2Loyola University Chicago, Chicago, Illinois, USA; 3 Aetna Foundation, Hartford, Connecticut, USA

Introduction

The Agency for Healthcare Research and Quality (AHRQ)-sponsored researchers identified ambulatory care sensitive conditions as conditions (ACSCs) for which hospitalizations can be prevented through primary care interventions. Prevention quality indicators (PQIs) measure the preventable hospitalizations for these conditions. Congestive heart failure accounts for a large proportion of hospitalizations due to chronic ACSCs. Affecting over 37 million individuals globally, congestive heart failure results in over a million hospitalizations annually.1,2,3

Methods

The state of Connecticut’s Acute Care Hospital Inpatient Discharge Database (HIDD) is maintained by the Connecticut Department of Public Health (CT DPH). Inpatient discharge data is submitted by acute care hospitals every six months to the Office of Health Care Access (OHCA)-approved vendor in January and July of each year. After clarifications with the source and addition of diagnostic codes, the vendor merges an annual file. The HIDD contains demographic, clinical, and billing data for discharges from all acute care hospitals in the state.4 This database was screened to obtain information regarding discharges for adult patients (>=18 years old) with the ICD-9 codes for congestive heart failure (PQI #8) during CMS fiscal year 2010 through 2012. The population rates of Prevention Quality Indicators (PQI) were calculated using the technical specifications from the AHRQ. We fit generalized estimating equations (GEE) to the 2012 data, in order to determine the effects of age, sex, race/ethnicity, payer status and median zip-code level income on the outcomes.

Results

A total of 26,894 CHF discharges were recorded during the study period. There was a net increase in the crude population rate for preventable hospitalizations due to congestive heart failure in Connecticut increased from 314.1 per 100,000 in 2010 to 322.6 per 100,000 in 2012. Upon stratification by sociodemographic variables, the rates were found to be higher for men, individuals older than 65 years, African Americans and Medicare enrollees. Multivariate analysis revealed that Medicaid enrollees (OR: 11.46; 95% CI: 10.25-12.82) had the highest odds of hospitalization due to CHF followed by Medicare enrollees (OR: 5.12; 95% CI: 4.67-5.61). African American (OR: 1.81; 95% CI: 1.68-1.94) and Hispanic (OR: 1.12; 95% CI: 1.02-1.22) race also demonstrated significantly increased odds. Men had 50% higher odds of hospitalization.

Conclusions

This study demonstrates the efficacy of an integrated state-level discharge database in monitoring ambulatory care sensitive conditions using a standardized approach. Over the long term, this tool could be used in identifying areas with low access to primary care and direct healthcare policies.

References

Improving quality of care for heart failure patients in a home care setting: what feedback information do nurses need?

Dawn W Dowding, PhD, RN1,2, Nicole Onorato, BS2, Jacqueline Merrill, PhD, RN1, David Russell, PhD2

1Columbia University School of Nursing, New York, NY; 2Visiting Nurse Service of New York, NY

Introduction

Providing performance feedback to clinicians is an important factor in improving health care quality1 and is a key component of the IHI Triple Aim initiative2. However it is unclear how best to deliver feedback to clinicians, especially in the home health setting. The objective of this study is to identify existing quality performance measures related to the care of patients with congestive heart failure (CHF) that are relevant to home care nurses, and that are actionable (i.e. under the control of the home care nurse), from the nurses’ viewpoint.

Methods

6 focus groups with home care nurses (N=61) working in a large not for profit home care agency in the Northeast region of the United States were conducted between November 2015 and February 2016. The number of groups was chosen to include nurses across all geographical regions of the agency. Focus groups were led by a member of the research team (DD or DR) with another member of the team taking notes (NO).

Focus group participants were provided with a list of 23 statements derived from existing evidence based practice guidelines on the management of patients with heart failure and what feedback on that item might contain (e.g. monitor weight and weight changes daily and notify MD if weight gain of 3lb overnight and over 5lb in 7 or fewer days. Feedback would provide indications of how weight has changed over time). Participants were asked to identify the top 5 statements that they felt were a priority in terms of receiving feedback and rank them from 1 (top priority) to 5 (least priority). The focus group discussion then explored the rationale for priority rankings and how a dashboard could be designed to provide that feedback. All focus group discussions were audio-taped and transcribed. Priority rankings were used to calculate a score (the higher the score, the greater the priority) for each statement. Qualitative data were analyzed using a framework approach, with codes assigned to the text inductively and deductively.

Results

Providing feedback to enable the tracking of vital signs, symptoms and weight changes was the highest ranked statement by nurses for receiving feedback (score =219). The second highest ranked statements related to providing feedback on what educational goals the patient had achieved (score = 75). Analysis from the focus group discussions identified that nurses wanted the feedback to be available in ‘real time’ when the opened the patient’s record in the agency EHR, and that it should be summarized in an accessible, visualized format (e.g. a line graph that displays trends over time).

Conclusion

This study has identified specific elements of care provision for heart failure patients that home care nurses would find valuable to receive feedback about, in order to improve care provision. These aspects of care are exclusively at individual patient level, and are required by nurses in real time. This is in contrast to more traditional approaches to providing feedback on care quality using dashboards, which are often aggregated data across patients (e.g. re-hospitalization rates) and rarely in real time, in a home care setting. The results of this study will be used to develop dashboards to provide feedback on quality outcomes to home care nurses, in real time, at the point of care.

References

Preliminary Evaluation of a Multi-Component Intervention Prior to Discharge to Reduce Pediatric 30-Day Readmissions

Amie J. Draper, MS1, Andrew Urbach, MD2, Roger Day, PhD1, Fereshteh Palmer, MS, RN2, Srinivasan Suresh, MD2, Fuchiang (Rich) Tsui, PhD1

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

Abstract: We performed a preliminary evaluation of a 4-component intervention (MAPS) to reduce 30-day readmission rates in a pediatric hospital. We distributed a survey to care providers and compared pre-and post-implementation readmission rates to assess impact on users, the discharge process and readmission rates. No definitive conclusions were made, but intervention improvements were identified. We plan to expand the evaluation by tracking compliance rates, examining user actions in the EHR, workflow analyses, and stakeholder interviews.

Introduction: On July 1, 2015, the Children’s Hospital of Pittsburgh (CHP) of UPMC, a 315-bed children’s general facility with 13,000+ annual admissions, completed deployment of a hospital-wide program called MAPS-to-discharge. The program helps ensure completion of specific pre-discharge care steps through formal, centralized documentation in a designated page within the electronic health record (EHR). The steps are organized into 4 components: 1) reconciling discharge medications (M), 2) scheduling future appointments (A), 3) completing patient education (P), and 4) providing appropriate social worker/care manager services (S). As prior studies have shown that similar multi-component interventions can improve 30-day readmission rates1, 2, we hypothesized that reduced 30-day readmission rates would be a by-product of implementing the program. In this study, we performed a preliminary evaluation of the MAPS program to determine its impact on users, the discharge process and 30-day readmission rates.

Methods: We performed both qualitative and quantitative evaluation. We sent out an 8-question survey to 1400 care providers within CHP. The survey was designed to get feedback on the following aspects of the MAPS documentation page: 1) usage and reasons for use, 2) quality of information, and 3) perceived impact on care processes and workflows. We then extracted month-level data from all in-patient EHRs for visits to CHP during 2012-2015. We compared 30-day readmission rates from July-December in 2015 to aggregate and individual rates from 2012-2014 over the same time period.

Results: Our survey response rate was 17%, with only 37% of respondents reporting that they use the MAPS documentation page. The most common use of the page was to prepare/review patient discharge instructions (81% of users), while the least common use was to take steps to prevent patient readmission (11% of users). Although 48% of users reported valuable information displayed on the MAPS page, few users reported it having an impact on care decisions or readmission risk assessment (16% and 11% of users, respectively). From open-response questions, we learned that users currently perceive an increased workload due to redundant documentation between systems during the transition to MAPS and missing information on the MAPS page. We also learned that not all users understand how to best utilize the MAPS page according to their clinical position. Readmission rates from July-December in 2015 were borderline significantly better than aggregate rates from 2012-2014 during the same months (p-value = 0.01); however, we observed no significant differences and borderline significant differences when comparing to individual rates for 2012, 2013, and 2014 (p-values = 0.63, 0.43, 0.01 respectively).

Discussion and Conclusion: Based on our preliminary findings, we could not make any definitive conclusions regarding the MAPS-to-discharge program. The findings indicate that the intervention may be valuable for the discharge process and might impact 30-day readmissions rates, but further analysis is required to confirm these findings. We plan to continue our evaluation by (1) correlating MAPS compliance rates (i.e. completion of each intervention component) to readmission, (2) correlating user actions within the EHR (via user log files) to readmission, (3) performing workflow analyses to understand user interactions with MAPS, and (4) perform interviews with care providers and patients (families). We will perform these evaluations after implementing the following recommended improvements: (1) promotion/education of MAPS program use for specific clinical roles, (2) embedded readmission risk score from a predictive model, and (3) indicator and alert system for MAPS component completion for each visit.

References
Evaluation of Clinical Score Usefulness to Guide Development of Automated Score Calculation for Hospitalized Patients

Mikhail A. Dziadzko, MD, PhD, Christopher A. Aakre, MD, Vitaly Herasevich, MD, PhD
Mayo Clinic, Rochester, MN

Abstract

Hundreds of clinical scores have been created, but relatively few are useful to clinicians in their routine clinical practice. We aim to identify the clinical scores deemed useful by hospital-based clinicians to prioritize development of automated scoring algorithms (calculators) within the electronic medical record.

Introduction

Clinical scores have been created to establish disease severity, to predict risk of specific outcomes, to estimate diagnostic probabilities, to be used in research settings, and to aid in clinical decision making. Of the scores that have been externally validated, it is unclear how many are clinically useful in practice. Many scores have been translated into clinical calculators and now are available online. These calculators need manual input – requiring clinician time for data collection and manual data entry. Therefore, automated score calculation within the electronic medical record provides an opportunity to save clinician time. To determine which clinical scores should be targeted for automated score calculation, we performed an exploratory multispecialty survey of hospital-based clinicians to determine the usefulness of evidence-based clinical scores.

Methods

109 readily available clinical calculators based on externally validated clinical scores for hospitalized patients were identified from the medical literature and online medical calculator repositories. Calculators were mapped to clinical specialties based on the age of the target population and the disease of interest. A web-based survey was created and distributed to a convenience sample of attending physicians at Mayo Clinic in Rochester, Minnesota. Respondents received surveys with calculators tailored to their specialty and were asked to rank each score on a 3-point scale: “Don’t need”, “Nice to have”, and “Very important”. Responses were weighted as follows to generate a proportion of clinicians that deemed each score useful:

\[ P = \frac{(\# \text{Very important}) + 0.5 \times (\# \text{Nice to have})}{(\# \text{Very important}) + (\# \text{Nice to have}) + (\# \text{Do not need})} \]

Scores with a proportion > 0.5 were classified as useful. Scores with 2 or fewer responses were not included in the final analysis.

Results

23 clinicians responded to the survey with specialties as follows: 15 anesthesia/critical care, 7 pediatric, 3 general surgery, 3 emergency medicine, and 7 internal medicine and subspecialties. 11 providers had multiple affiliations. 15 calculators were excluded (2 or fewer responses). 25/109 (22.3%) calculators were identified as useful (Table 1).

Table 1. List of useful clinical calculators identified by respondents.

<table>
<thead>
<tr>
<th>CHADS2 Score</th>
<th>PESI score</th>
<th>Simplified PESI score</th>
<th>SIRS/Sepsis/Septic Shock Criteria</th>
<th>ASA Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nexus Criteria for C-spine imaging</td>
<td>PERC rule</td>
<td>Well’s Criteria for DVT</td>
<td>Well’s Criteria for PE</td>
<td>PELD score</td>
</tr>
<tr>
<td>Clinical Pulmonary Infection Score</td>
<td>LRINEC score</td>
<td>Oxygenation index</td>
<td>SOFA score</td>
<td>Alvarado score for acute appendicitis</td>
</tr>
<tr>
<td>Canadian CT head rule</td>
<td>Ranson’s criteria</td>
<td>PSI/PORT score</td>
<td>MOD score</td>
<td>APACHE II score</td>
</tr>
<tr>
<td>PASS score</td>
<td>SAP II score</td>
<td>4Ts Score for HIT</td>
<td>FOUR score</td>
<td>CURB-65</td>
</tr>
</tbody>
</table>

Conclusion

Fewer than one quarter of clinical calculators based on externally validated clinical scores were classified as useful in our analysis. Most of the scores classified as useful in Table 1 are used for diagnosis and prognosis of acute illness, which may reflect the anesthesia/critical care or internal medicine backgrounds of the majority of respondents. Our findings are consistent with previous observations on the rejection of prognostic models by clinicians(1), but should be replicated in a multicenter study before committing resources to automate calculation of the most useful scores. Similar multicenter studies in ambulatory settings may be useful to determine useful scores in outpatient practice.

References

Analyzing Electronic Health Record Interactions to Capture Resident Work Hours

Adam C. Dziorny, MD, PhD; Evan Orenstein, MD; Robert B. Lindell, MD; Nicole A. Hames, MD; Bimal Desai, MD, MBI
Department of Pediatrics, The Children’s Hospital of Philadelphia, Philadelphia PA

Introduction: Recording and monitoring of resident duty hours is a program mandate from the ACGME to prevent resident burnout and avoid medical errors. Annual review of rotations with duty hour violations provides room for program improvement. Self-report of duty hours is onerous, frequently leading to under- or over-reporting and associated challenges with compliance. Residents constantly interact with the electronic health record (EHR) while on inpatient services. EHR access logs provide an opportunity to capture the minimum hours of resident work in a non-intrusive but quantitative manner. In our large academic medical center with approximately 145 pediatric residents, our EHR, EPIC, stores time stamps of interaction events (e.g. open list, view results, sign order) for all users. We have developed an algorithm to convert these time stamps into shift start and end times.

Objectives:
- To characterize frequency of events, and intervals between the EHR time stamps for residents of different training levels working in differing inpatient roles.
- To compare EHR time stamp generated shift estimates to scheduled resident shifts.

Methods: For the purposes of analyzing EHR interactions, pediatric residents were characterized into one of three inpatient roles: intern; “resitern” or senior resident acting as a front line ordering clinician; and senior supervising resident. EHR time stamps were retrieved from the EHR database system, matched to scheduled resident shifts, de-identified and analyzed. Time stamp intervals between events were computed and binned in Excel. Assumptions of continuous shifts were made based on the measured time stamp intervals, and shift times were estimated from raw time stamp data using custom software. Estimated shift times were compared to scheduled shift times and percent agreement was calculated as overlapping hours divided by total hour span.

Results: On a per-day basis, interns and “resiterns” interacted the most with the EHR, generating the most logged events [Interns: 3153 (SD 876); Resiterns: 3434 (SD 764) interaction events per day] compared to our senior supervising residents [1938 (SD 35) events per day]. Time interval histograms demonstrated that among interns, 100% of the time intervals between events were either < 120 minutes or > 8 hours. Intervals among “resiterns” showed 0.001% (1 interval out of approximately 96000) between 120 – 180 minutes with the remainder either < 120 minutes or > 8 hours. Senior supervising residents demonstrated multiple intervals between 120 – 480 minutes (6 total intervals out of ~ 27000) and required different cutoffs when calculating shift times. Estimate shifts overlapped with scheduled shifts 91% (SD 6%) among interns, 87% (SD 12%) among “resiterns” and 91% (SD 5%) among supervising seniors. Patterns of non-overlap include “resitern” estimated shifts consistently ending earlier than scheduled (mean: 53 minutes earlier) and senior shifts starting later than scheduled (mean: 14.8 minutes later).

Conclusions: Residents in a large academic children’s hospital interact frequently with our electronic health record. Those residents providing front line care to patients interact more frequently than supervising residents. From our sample data, in a typical shift, front line ordering residents are rarely away from the EHR for more than 1 hour, and never away for more than 3 hours. Shift times estimated from EHR access logs demonstrate good concordance with scheduled shift times for all trainee roles. Patterns suggest that “resiterns” stop interacting with the EHR earlier than scheduled, and senior supervising residents do not interact with the EHR as early as expected in a given shift. These differences may reflect the actual shift times the residents worked, as is likely in the case of “resiterns” who typically leave when work has been completed for the day. However the differences also may reflect that not all resident workflow requires EHR interaction, as is likely the case with senior supervising residents.

Accurate shift times generated by interaction with EHR systems provide useful, quantitative means to assess work hours. Future work involves using estimated shift times to determine duty hour adherence based on ACGME requirements and rotation-based analysis to look for areas of residency program improvement.
An Analysis of Readmission Events over Time in Patients with Type 2 Diabetes
Dara L. Eckerle Mize, MD1,2, Mia Levy MD, PhD1,3, Shubhada Jagasia MD, MMHC2, Colin Walsh, MD, MA1,4
1Department of Biomedical Informatics, 2Division of Diabetes, Endocrinology & Metabolism; 3Division of Hematology & Oncology; 4Division of Internal Medicine, Vanderbilt University Medical Center, Nashville, TN

Introduction
Financial penalties have been imposed on healthcare systems which fail to meet accepted quality standards for excess hospital readmissions. Patients with diabetes account for 25-30% of the hospitalized population. Diabetes is even more common among the most common conditions requiring readmission within 30-days of hospital discharge. The best approach for implementing safe care to diabetic patients across the continuum remains unknown. We hypothesize that trends in readmission diagnosis change over time in patients with type 2 diabetes. Identified variation in readmission diagnosis can be used to improve the utility of readmission risk prediction models when compared to traditional all-cause, 30-day readmission risk prediction models.

Methods
Subjects included all adults admitted to Vanderbilt University Medical Center (VUMC) between October 1, 2010 and September 15, 2015 with a diagnosis of T2DM at the time of hospital discharge and >1 admission during the study period. We obtained data from the Vanderbilt Research Derivative (RD), a database of clinical and related data derived from VUMC’s clinical systems. Descriptive statistics were used to summarize the discharge and readmission rates of patients with T2DM and to analyze variation in readmission diagnosis of patients with T2DM over time. Data visualization techniques temporally stratified trends in the variation in readmission diagnosis.

Results
As shown in figure 1, distribution of readmission diagnosis varies within each time period. Over time, the number of readmissions decrease but these preliminary results do not support a clear cutoff at 30 days as is currently used to judge quality over the continuum of care. Additionally, readmission diagnoses do not change consistently over time. For example, figure 2 demonstrates that congestive heart failure (CHF) is the most common reason for readmission in every time period except the first week after discharge. A larger dataset demonstrates that CHF accounts for 20-25% of readmissions and remains the most common readmission diagnosis 12 months after discharge. In contrast, renal failure and infection appear to account for a greater proportion of readmissions the first few weeks after discharge and decrease over time. Statistical analysis on a larger dataset is underway.

Conclusions
Using a data-driven approach to the development of a hospital readmission risk prediction model, we identified variation in the readmission diagnoses of diabetic patients over time. Readmission risk prediction models which take readmission diagnosis into consideration may outperform traditional all-cause models and inform the allocation of resources to reduce readmission risk.
A Maximum Entropy Model of Follow-up: Using Cancer Registry and Insurance Data to Identify Encounters for Breast Cancer Surveillance

Kathryn S. Egan, MS¹, Gary H. Lyman, MD, MPH¹,², Karma L. Kreizenbeck¹, Catherine R. Fedorenko¹, April Alfiler¹, Heather Noble¹, Tracy Kusniir-Wong³, Ada Mohedano³, F. Marc Stewart, MD²,³, Benjamin E. Greer²,³, Scott D. Ramsey, MD, PhD¹,²
¹Fred Hutchinson Cancer Research Center, Seattle, WA; ²University of Washington, Seattle, WA; ³Seattle Cancer Care Alliance, Seattle, WA

Introduction

National guidelines recommend that women who have undergone curative treatment for breast cancer do not receive surveillance testing¹. We have begun collecting data at a regional oncology clinic to measure adherence to this recommendation in the first year of surveillance. However, because follow-up visits for breast cancer are not specified in insurance data or the cancer registry, determining if and when a patient undergoes surveillance has required intense manual abstraction of clinical records. To reduce the need for abstraction, we developed a statistical classifier that identifies follow-up visits for eligible patients with 88% sensitivity and 91% specificity. The classifier identifies patients in their first year of surveillance with 100% sensitivity and 86% specificity.

Methodology

We extracted insurance billing data at a regional oncology clinic for patients with one or more clinic visits in 1/1/2012-12/31/2014 with a code for breast cancer and no code for treatment (4203 patients). We used the clinic’s cancer registry to filter for breast cancer patients who had stage I-IIIA disease, no prior history of cancer, no concurrent non-breast cancer, and had undergone treatment with curative intent (584 patients). Dates of end of treatment and routine breast cancer follow-up were abstracted for the first 200 patients, alphabetical by medical record number. We trained a maximum entropy classifier on 180 randomly selected patients, one vector per discharge date per patient. Twenty patients were held out for testing. Vectors were uniformly weighted and consisted of a boolean indicating whether the encounter was pre- or post-end of treatment, and all billing codes for that encounter, labeled by type (e.g. first ICD-9 diagnosis code). End of treatment dates for the model were calculated using the cancer registry: surgery date, chemotherapy start date + 90 days, or radiation start date + 30 days, whichever came last (90 and 30 days produced optimal sensitivity to follow-up visits in training).

Results

Against abstracted data, classifier sensitivity and specificity for individual encounters were 88.2% and 91.0%, respectively (Table 1). Out of the 20 test patients, 13 experienced routine follow-up for breast cancer in the first 13 months after end of treatment. The model identified all 13 of these patients with only one false positive (Table 2).

Discussion

The classifier has several limitations: abstracting alphabetically may have introduced bias; we excluded patients receiving initial diagnosis or treatment elsewhere and transferring to the current clinic for surveillance; the model is untested at other clinics (although regulation of registry and insurance data suggests it is transferrable); small sample size; and abstraction of breast cancer follow-up visits was subject to interpretation. Nonetheless, the classifier can play a crucial role in quickly and accurately evaluating clinic-wide adherence to national guidelines.

References


Table 1. Results by encounter.

<table>
<thead>
<tr>
<th>Abstracted</th>
<th>Follow-up</th>
<th>Not follow-up</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up</td>
<td>119</td>
<td>16</td>
<td>135</td>
</tr>
<tr>
<td>Not follow-up</td>
<td>31</td>
<td>312</td>
<td>343</td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>328</td>
<td>478</td>
</tr>
</tbody>
</table>

Table 2. Results by patient.

<table>
<thead>
<tr>
<th>Abstracted</th>
<th>Follow-up</th>
<th>Not follow-up</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up</td>
<td>13</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Not follow-up</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>6</td>
<td>20</td>
</tr>
</tbody>
</table>
Identifying SNOMED Concepts Relevant to CHA₂DS₂–VASc and HAS-BLED Scores¹

Peter L. Elkin, MD, Edwin Anand, MD, Chris Crowner, MS, Sina Erfani, BS, Grégoire Ficheur, PhD, MD, and Daniel R. Schlegel, PhD
Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA

Introduction

The CHA₂DS₂–VASc¹ and HAS-BLED² scores are used to assess the risk of stroke in patients with atrial fibrillation, and risk of major bleeding due to oral anticoagulants used to treat atrial fibrillation, respectively. Our eventual goal is to build a reasoning model to calculate these scores automatically. In this study we identify a list of SNOMED concepts which are relevant to the generation of these two scores.

Methods

Two clinicians, EA and GF, familiar with CHA₂DS₂–VASc and HAS-BLED used the IHTSDO SNOMED CT Browser to find concepts relevant to the clinical use of those scales. The clinicians examined all SNOMED concepts which were found when searching for terms the clinicians thought were relevant to each criteria in the two scales. They examined all descendants of relevant concepts to ensure they are also relevant, and noted if they were not. A consensus³ was formed with the help of an adjudicator (PLE). A Cohen’s κ score could not be calculated because the (potentially very large) set of SNOMED concepts which both clinicians found to be irrelevant is unknown, and an estimate could yield an artificially high κ score. Instead we report the further three variables in the contingency table. We computed these three values both at the top level (the one selected by the clinicians) and including descendants.

Results

Of the 106 concepts identified by the clinicians, only 13 (12%) of them were common to both. Including the descendants the intersection was just over 47%. The variability in depth of the tree under selected terms can be seen in Figure 1 – GF selected 36 concepts which EA did not, but including descendants only added 42 concepts to the total, while the 57 which EA selected that GF did not added 5,862. The final consensus included 47 concepts, indicating a rather large amount of culling during arbitration was performed.

Discussion

The overwhelmingly majority of issued resolved during arbitration was the removal of concepts for which an ancestor had already been identified. This accounts for 36 of the 59 concepts excluded during arbitration. In other cases a category was chosen which was too broad – this happened twice and accounts for many of the descendants EA captured but GF did not. Other issues resolved during arbitration had to do with whether to include items which have low probability of indicating the criteria, e.g., Endocrinologist for renal disease or Respiratory crackles for congestive heart failure. It was decided that these should be excluded. Finally, some clinical decisions had to be made. For example, it was decided during arbitration to include Chronic liver disease instead of Disorder of liver, which one clinician had identified, since the scales are about longer term effects, rather than short-lived diseases.

References


¹ This study is sponsored by Pfizer.
² The final concept list is available on Google Drive at http://bit.ly/1U4kmpe
A Bespoke Big Data/Visual Analytics Solution for Treatment Pathway Analysis and Visualization Using Healthcare Data

Sergio Eslava, MD, MMI¹, John Cai, MD, PhD¹, Saar Golde, PhD², Fei Yang²
¹Celgene Corporation, Summit, NJ, ²Knowledgent, Warren, NJ

Introduction

Multiple areas of drug development require a detailed understanding of how drugs are prescribed by healthcare providers and used by patients outside the controlled environments of clinical trials, and for a long time the industry has used healthcare databases (EMR, Claims, Registries, etc) for that purpose.

However, to extract drug utilization patterns from those data sources and present them in an intuitive manner that facilitates decision making, standard analytical methods need to be supplemented with Visual Analytics¹.

Methods and Results

We custom built an interactive Visual Analytics solution using only Open Source analytical tools (R, RStudio, Shiny and D3.js) on top of a Hadoop/AWS cluster, with the following functionality:

- Ingest patient level data from an administrative claims database (Truven’s Marketscan) and build disease-specific patient cohorts from standardized phenotype definitions
- Extract sequential events (diagnoses, prescriptions, procedures, providers) and reconstructs treatment pathways from those events using pre-defined business rules
- Present those patterns to end users using customized data visualizations at the population level (lines of therapy visualization using Sankey diagrams²,³) and at the patient level (individual patient timelines) (Fig 1)
- Summarize and display patient characteristics (demographics, comorbidities), cost and healthcare resource utilization, duration of treatment and time to next treatment in relation to different pathways

![Figure 1. Left: Line of therapy visualization for Chron’s Disease using Sankey Diagrams. Right: Individual Patient Timeline for a Chrohn’s Disease patient](image)

Discussion

Our solution has been used in diverse disease areas including Crohn’s disease, AML and Psoriasis, where it has proven effective in helping to understand the patterns of care for a disease-specific patient population.

Future directions include the incorporation of additional data sources including EHR and integrated EHR/Claims data to help overcome some of the limitations of using claim databases alone.

References

A Multi-Institutional Honest Broker in the Cloud
Claudiu Farcas, PhD, Tyler Bath, Paulina Paul, Antonios Koures, PhD
Lucila Ohno-Machado, PhD, MD, MBA
University of California, San Diego

Abstract
The cloud paradigm is permeating scientific advancements, yet it brings its own challenges for human-subjects research. We introduce a novel honest broker system to aid investigators in safe management and exchange of Protected Health Information (PHI). Our approach takes advantage of the private iDASH HIPAA CLOUD and its associated PHI repository. Our system integrates institutional policies that govern data exchange.

Introduction
Current research practices involving data from multiple research institutions depend on human-in-the-loop processes for creating the appropriate request forms for each institution, specifying the data of interest, performing the extracts from the available EHR systems, transferring the data to the institution originating the request, etc. Some of these processes can be highly automated, others require intensive human intervention (e.g., IRB reviews and approvals). To address these challenges, we developed a prototype system that acts as a broker to mediate the end-to-end processes, including the physical delivery of data, and keeps the researcher and relevant staff from several institutions in a continuous communication loop with regard to each study and their progress. It leverages the private IDASH HIPAA Cloud, developed as part of the NIH NCBC program, to store sensitive data in a secure isolated repository on behalf of the requesting party and institutions involved in each study.

The Honest Broker
Our prototype matches (1) a technological solution to safely request, extract, aggregate and compute on human-subjects data in a secure environment, with (2) a legal and organizational framework, including handling of data-use agreements and IRB approvals across institutions. It gives participating institutions full control over who has access to the data and for what purpose. At the same time, it aims at reducing the administrative and technical burden on investigators and supporting staff. The legal framework encompasses the handling of data sharing agreements between relevant parties, with support for de-identified, limited, or fully identifiable information, along with specifications on authorized use by investigators. The system translates these restrictions into computer actionable items, such as access rights onto specific repositories, folders, and files, and automates the collection of data request parameters, notification of the appropriate staff at each institution participating in a study, estimated completion dates and progress reporting, and data transfers to a secure repository.

This work addresses some of the Findability + Accessibility + Interoperability + Reuse (FAIR) aspects of biomedical research that are currently hampered by siloed data, unsafe computing practices, inadequate tools and poor documentation

![Legal Framework Diagram](Image)

**Figure 1.** The Honest Broker combines a legal and technical solution to streamline the end-to-end data access
Interoperability among Prenatal EHRs: A Formal Ontology Approach

Fernanda Farinelli, D.S.I.S.1,2,3, Mauricio B. Almeida, PhD1, Peter Elkin, MD2, Barry Smith, PhD2,3

1 Federal University of Minas Gerais, Brazil; School of Information Science
2 Department of Biomedical Informatics and 3 Department of Philosophy, University at Buffalo, USA

Introduction

Gathering information from EHRs connected to different information systems is a challenge and involves the adoption of interoperability solutions(1). To overcome this interoperability failure among prenatal EHRs our strategy is to develop an ontology in the obstetric and neonatal domain (OntONeo(2,3)). Such ontology will be able to join different standards and terminologies adopted by information systems that deal with prenatal EHRs. OntONeo has the potential to contribute to interoperability of information among EHR from different specialties. In addition, the definitions of OntONeo will facilitate the understanding of how information can be organized in EHRs for purposes of healthcare.

Methodology

We adopted Basic Formal Ontology (BFO) as the top-level ontology of OntONeo, the methodology of ontological realism(4), and we are also following the Open Biomedical Ontologies (OBO) Foundry(5) principles.

Partial results

We built a formal definition to specialties found in EHRs on OntONeo domain (Figure 1). We identified a set of basic types of information that are common across different EHRs independently of specialty (e.g. demographic data and vitals). Moreover, each medical specialty has specific information about the care provided.

Figure 1. Part of OntONeo that deal with the basic information common in EHRs.

Conclusion

We presented how OntONeo ontology represents the EHR data involved in the care of pregnant women. This is an on-going project, and the current version can be found at http://ontoneo.wordpress.com.

Acknowledgments

We thank CAPES and CNPq for financial support to the primary author. This work is also supported in part by the NIH NCATS under CTSA award Number UL1TR001412.

References

Why is Natural Language Processing so Difficult to Generalize?

Jeffrey P Ferraro1,2, Peter J. Haug, MD1,2, Michael Wagner, MD, PhD3
1Intermountain Healthcare, Salt Lake City, Utah; 2University of Utah, Salt Lake City; 3University of Pittsburgh, Pittsburgh, Pennsylvania

Introduction
Clinical natural language processing (NLP) systems are becoming an integral component of systems that relay on information contained within unstructured and semi-structured clinical notes. These systems include disease surveillance systems and predictive diagnostic modeling systems to name a few1,2. Some of these systems such as disease surveillance require portability and ease of deployment across geographical regions and include natural language processing tasks for critical information mining.

The information contained within unstructured clinical notes takes many forms and differs in content and structure within institutions as well as across institutions. To some degree, natural language processing tools must be resilient to these local differences if they are to be effective in providing critical information while supporting system portability.

Methods
As part of a natural language processing compatibility study to support disease surveillance we evaluated some of the difficulties that arise in developing a generalizable NLP solution. The goal was the extraction of 70 clinical findings associated with Influenza at two large healthcare institutions in North America. We analyze ED encounter clinical notes at each institution to determine the distributional characteristics of the language used to describe each of these 70 clinical findings. For example, for the clinical finding “chest wall retractions - present” one institution may describe it as “was using his accessory muscles for respiration” while the other institution may describe it as “does have some abdominal breathing and moderate subcostal and mild to moderate suprasternal retractions”. We confirm the statistical validity of distributional differences between institutional languages by applying Fisher’s Exact Test for Homogeneity performed on the word frequency describing each clinical finding.

Results
P-value significance was adjusted for multiple comparisons testing using the false discovery rate method. Seventy percent (n = 49/70) of the clinical findings had statistically significant (adjusted p-value < 0.05) differences in language used to describe the clinical findings between the institutions.

Discussion
It is common to find language differences in clinical notes describing clinical findings among institutions. Even within institutions, dictation styles and linguistic expression may vary among clinicians. This is one of the most difficult challenges faced in generalizing modern NLP systems today. Whether rule-based or statistically based, NLP systems are developed from training sets providing samples of phrases and linguistic expression to draw upon in developing extraction rules or statistical extraction methods. Systems that show reasonable generalizing characteristics have typically done a good job at incorporating unseen-case heuristics into their extraction algorithms in anticipation of variance seen in practical application. In information extraction, it is challenging to develop good unseen-case guessing heuristics because there is typically over-fitting caused by the institution-dependent lexical scope of relevant phrases. This applies to the phrasal components used to express clinical concepts as well as to the extraction rules derived to recognize them. It is also difficult to develop general extraction methods to address synonymy at a complex phrasal level.

These are some of the reasons why it is so difficult to generalize natural language processing tasks such as information extraction. Our poster presentation will elaborate on these challenges, provide illustrative examples of typical, location-dependent, textual variation, and discuss existing and proposed approaches to dealing with this ubiquitous problem.

References
Review and Evaluation of the State of Standardization of Computable Phenotype
Stephanie Feudjio Feupe, MSc, Ko-Wei Lin, PhD, Tsung-Ting Kuo, PhD, Chun-Nan Hsu, PhD, Hyeon-Eui Kim, PhD
Department of Biomedical Informatics, UC San Diego, La Jolla, USA

Abstract
Inadequate standardization of phenotype data in biomedical database and electronic medical record (EMR) has created challenges in data reuse for patient care, quality improvement, new scientific discovery, and validation of existing knowledge. We conducted literature review on this topic published in the past 10 years to assess the efforts on phenotype standardization and rendering phenotype computable. We summarize the findings here.

Introduction
According to the medical encyclopedia, a phenotype is an observable or measurable characteristic or trait in an organism. In the clinical domain a phenotype can be a disease or condition, a symptom, a biomedical process or a measure of physical characteristic. The ever growing need to investigate the associations among phenotypic data and other types of data such as genetic, pharmaco-genetic, environmental, psycho-behavioral data makes it crucial to have standard for phenotype representation across databases. While the availability of phenotype data in the EMR, clinical databases and other supporting sources bring new opportunities for large scale data driven research such as genome wide association studies and cohort discovery; the lack of clarity in phenotype definitions and the idiosyncrasies in data representations make it a non-trivial task to render phenotype data interoperable and sharable. Given the significance of phenotypic data and the growing efforts to make phenotype data computable, we reviewed and analyzed published literature to access current state of the research in order to understand what efforts have been done, to identify potential gap and unexplored areas for developing further research focus on this topic.

Methods
We searched articles from both PubMed and Web of Science (WoS) using the following queries respectively.
- PubMed: Phenotype[MeSH] AND (standardization OR "vocabulary, controlled"[MeSH] OR "terminology as topic"[MeSH]) AND Published year: past 10 years AND Language: English AND Abstract available AND Human studies
- WoS: TOPIC: (Phenotype OR "computable phenotype") AND TOPIC: ("standardization, definitions" OR "vocabulary, controlled" OR "terminology as topic") AND YEAR PUBLISHED: (2004-2015) AND Language: English

The searches yielded 611 and 440 papers respectively; a total of 918 papers after duplicate removal. Using our predefined inclusion and exclusion criteria, we selected 48 articles. To be included, articles had to focus on phenotype standardization with an informatics-based method for standardization and suggest or present some evaluation of their approaches. We further screened the 48 articles by reading the full-text papers. Articles related to establishing clinical guidelines, papers that dealt with diagnostic criteria as opposed to phenotype description were removed. Fifteen relevant articles met our eligibility criteria, thus included in our review. The articles were summarized using a standardized review form, which is designed to capture study characteristics such as study author, year, phenotype target domain, study goal, data source, research methods, results, open sources, evaluation methods, and limitations of each paper.

Results and conclusion
Phenotypes are considered computable and integrable when they can be described to a more granular level as possible. Methods for such representation ranged from annotation by human, to standardization based on computerized from structured and unstructured data sources. In the past year, efforts like eMERGE, PhenX, HPO have attempted to standardize phenotype data from omics research databases; whereas, the combination of information models and terminology system like UMLS and SNOMED-CT has been used to assure EHR interoperability and phenotype identification. While most methods for rendering phenotype computable in these studies seemed to be informative, the majority of the work we reviewed focused on specific types of phenotype or the data from a single data repository, which can be considered a potential threat to generalizability.

Acknowledgement
This study was supported in part by NIH/NLM training grant (T15 LM011271), PCORI grant (CDRN-1306-04819), and NIH/NHLBI grant (U54HL108460).
Number of patients included in randomized controlled trials versus observational studies in three reference journals over 20 years

Grégoire Ficheur MD-PhD, Daniel R. Schlegel PhD, and Peter L. Elkin MD
Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA

Introduction
Empirically, many informaticians have noted an increase in clinical research networks and data warehouses which have as part of their purpose to serve observational research needs. It is unknown whether this empiric trend has led to an increase in the size of studies in terms of number of participants recruited in trials or in observational studies. Our objective was to estimate the number of patients included in randomized controlled trials versus observational studies in three reference journals from 1995 to 2014.

Methods
The structured abstracts from the New England Journal of Medicine, the Lancet and the Journal of the American Medical Association we extracted from MEDLINE for the period 1995-2014. Abstracts including a Methods section were retained. Two lists of patterns were built by an expert in order to automatically analyze the Methods section then classify the abstracts into the two categories randomized controlled trials (RCTs) or observational studies (OS). A list of regular expressions was then used to automatically extract the number of participants included in each study from the Methods section. Four successive periods of 5 years were defined and the median was computed for each of these periods. The corresponding boxplots were constructed without the potential outliers.

Results
The analysis included 13,377 abstracts. For the four successive periods of interest, the medians of the number of participants were respectively 289, 436, 668 and 512 for the RCTs. For the same periods, the medians were respectively 446, 558, 1338 and 2067 for the observational studies. The corresponding boxplots are presented in Figure 1.

Discussion
Contrary to RCTs, the number of patients included in observational studies has increased strongly over the last 20 years. The move towards dematerialized electronic medical records and the reuse of this data for epidemiology seems a reasonable hypothesis to explain this observation.
Using Machine Learning to Improve Sentence Detection

Dezon K. Finch, PhD\textsuperscript{1}, Stephen L. Luther, PhD\textsuperscript{2}
\textsuperscript{1}HSR&D Center of Innovation in Disability and Rehabilitation Research.
James A Haley Veterans Hospital, Tampa, FL

Introduction

Medical progress notes in the Veterans Health Administration have proven to be challenging for standard sentence splitters due to the presence of semi-structured data and because free text is often not written in grammatically correct sentences. This makes it difficult for Natural Language Process systems to accurately extract concepts from the text. Sentence splitters either split sentences apart because of the new line characters (\texttt{\textbackslash n}) left in the text or they find it very difficult to identify the ends of sentences because punctuation is absent.

Methods

100 progress notes were sampled from an existing study corpus such that each note was unique based on the standardized title. A Python script was developed to sample text strings in two ways. First, the samples were taken with the new line character at the center with 10 characters on each side and a large span of 20 characters on each side. Second, samples were taken such that characters that are expected at the end of sentences were at the center, such as periods, question marks, exclamation points and quotations. Each character of each string was coded with the character type as the value of a variable that indicated its position in the string. Each string with the new line character in the center was hand labeled as True when break was within the boundary of a sentence and False when it was not. Refer to Table 1 for illustrative examples. The `|` denotes `\texttt{\textbackslash n}` character in the text.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Target</th>
<th>B8</th>
<th>B7</th>
<th>B6</th>
<th>B5</th>
<th>B4</th>
<th>B3</th>
<th>B2</th>
<th>B1</th>
<th>C</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Join candidate</td>
<td>u c k w a s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>\texttt{\textbackslash n}</td>
<td>h a u</td>
<td>l i n</td>
<td>g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coded True</td>
<td>lo</td>
<td>lo</td>
<td>lo</td>
<td>sp</td>
<td>lo</td>
<td>lo</td>
<td>sp</td>
<td>nl</td>
<td>lo</td>
<td>lo</td>
<td>lo</td>
<td>lo</td>
<td>lo</td>
<td>sp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Split candidate</td>
<td>E D</td>
<td>B Y</td>
<td>P</td>
<td>T</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coded False</td>
<td>up</td>
<td>up</td>
<td>sp</td>
<td>up</td>
<td>up</td>
<td>sp</td>
<td>up</td>
<td>up</td>
<td>pp</td>
<td>sp</td>
<td>nl</td>
<td>sp</td>
<td>sp</td>
<td>sp</td>
<td>sp</td>
<td>sp</td>
<td>up</td>
<td>lo</td>
</tr>
</tbody>
</table>

The strings with end of sentence characters were labeled True when it was a sentence boundary and False when it was not. The 20 character strings were left in their raw text form for the labeler to reference. Two separate machine models were developed using C5 rule induction and 10 by 10 cross validation. The trained models are converted to executable code to be used in a pre-processing program that would extract the strings from a note, code the characters in the string, determine if a join should be made by changing the `\texttt{\textbackslash n}` to a space and whether candidate sentence ends should be split, by inserting a changing a space to a `\texttt{\textbackslash n}`.

Results

The model developed for identifying sentence splits (n=18,600) achieved an overall accuracy of 98.5% while the model for identifying joins (n=48,900) achieved an accuracy of 98.9%. The split detection model was able to distinguish recognize typical abbreviations like “Dr.” and in cases where sentences ended with quotation marks, split on the quotation rather than the period. There were some cases where text was written in all caps that were missed. The joins classifier accurately identified cases where some text was indented and overcame multiple leading spaces in the gap. However, in cases where text was deeply indented it was not able to properly identify the break in sentences due to the limitation of the length of the string sample. Some of this was overcome by reducing the amount of leading white space before extracting the string samples.

Conclusion

The tool produces note with each sentence complete and on a single line which will make parsing algorithms in NLP perform more effectively. We will test the effect on parsing accuracy in future efforts and make the final version in Java available to the public for use NLP pipelines. More complete examples provided in the poster.
Informatics Challenges in Working with "Big Data" – A Use Case in Identifying Predictors of Constipation

Fern FitzHenry1,2 Svetlana K. Eden1,2 Jason Denton1,2 Robert J. LoCasale3 Hui Cao3 Aize Cao1,2 Ruth Reeves1,2 Nancy Wells1,2 Michael E. Matheny1,2

1Tennessee Valley Healthcare System, Veterans Affairs Medical Center, Nashville, TN; 2Vanderbilt University Medical Center, Nashville, TN; 3AstraZeneca Pharmaceuticals, Gaithersburg, MD

Introduction

Large datasets just by virtue of their size can present problems in data convergence. The electronic health records of the Veterans Health Administration (VHA) presents a rich use case of a large dataset, studying the risk factors for developing constipation in opioid naïve veterans. When dealing with large observation cohort data, study design is complicated by computational resources required and model convergence challenges. We sought to describe some of the analytic challenges in the use of “Big Data” within the use case of opioid-associated constipation.

Methods

We selected a national cohort of veterans receiving care from 1-2008 to 5-2013 with opioids for 30 days or more (any formulation/dose) and at least one encounter in each of the previous 2 years, but with no opioid exposure or cancer diagnoses in the year prior. The analysis was a Cox Regression studying the association of subject characteristics measured in the year prior to the opioid exposure with the development of constipation. The outcome, time to constipation, was defined as time to constipation related diagnosis, procedure, or new/changed laxative treatment. Out of 162 available variables, 82 variables with a prevalence <1% or >99% were excluded to allow convergence of the model. The final model adjusted for the remaining 80 variables. The analysis was performed in statistical language R.1

Results

The number of veterans meeting the criteria for inclusion/exclusion was 861,422 with 11.42% developing constipation. The top 14 significant variables (7 positive and 7 negative) associated with the development of constipation are presented in Figure 1. Although the number of events (98,360) was large enough to include all 162 variables, some variables had to be excluded because the full model would not converge. After identifying and removing variables causing computational singularity, the model still produced an aberrant 3 digit effect size. Further investigation showed that the reason for singularity and the aberrant effect size was that some binary variables were very sparse (for example, a rare medication taken by a single patient who developed constipation). Removing sparse variables with a prevalence <1% or >99%, an a priori defined threshold, solved the problem of convergence. The modeling showed clinically plausible associations with a C-index of 0.676 (95% CI: 0.676, 0.678).

Conclusion

A careful understanding of the data and methods are required to execute analyses in “Big Data." Exploring non-linear functional forms of association between the explanatory variables and the outcome might help improve discrimination ability of the model.

Acknowledgements: This study’s work was supported with resources and the use of facilities at the VHA Tennessee Valley Health System and was funded by AstraZeneca Pharmaceuticals, Gaithersburg, MD.

Counting Knowledge Objects – Estimating How Many Discrete Knowledge Artifacts are described in the Biomedical Literature by Type

Allen J. Flynn, PharmD1,2, Charles P. Friedman, PhD1,2,3
1School of Information, 2Medical School, 3School of Public Health, University of Michigan, Ann Arbor, MI

Abstract

Some knowledge artifacts are intended for routine use. These include calculators, scales, algorithms, checklists, flowcharts, guidelines, pathways, protocols, and templates. Bibliometric data from PubMed were used to estimate quantities of these types of “knowledge objects.” Since 1970, about 47,000 articles have described such objects.

Introduction

In a Learning Health System, publication of knowledge artifacts must be followed by their routine use.1 To understand the scope and scale of the knowledge management and implementation challenges that need to be overcome to achieve a Learning Health System, an estimate of the number of potentially useful knowledge objects is needed. This estimate informs ongoing work to develop a digital library system within which to represent, store, curate, manage, and make accessible biomedical knowledge artifacts as digital objects.

Method

Nine coordinated PubMed database article title queries were performed and analyzed using NCBI’s Entrez E-utilities2 and PERL (v.5.12.4). First, a query for ‘calculator’ was run and every article found was manually reviewed so that all clinical calculators could be counted. Second, a Boolean logic query for ‘scale’ and (‘clinical’ or ‘treatment’ or ‘management’ or ‘practice’) was run with keyword exclusions (e.g., ‘pay’). Finally, seven more queries were run to find and count relevant article titles with ‘algorithm’, ‘checklist’, ‘flowchart’, ‘guideline’, ‘pathway’, ‘protocol’, and ‘template.’ To avoid duplicates, each successive query excluded all previously queried artifact types. PERL scripts organized and counted results. Results are limited to publications indexed in PubMed.

Results

From 1970 through 2015, guidelines are featured in 27,268, protocols in 9,652 and algorithms in 3,552 articles, respectively. Taken together, another six knowledge artifact types are featured in 6,342 articles. Accounting for some duplication but also allowing that some articles discuss more than one artifact, it is estimated that since 1970 about 47,000 articles in PubMed describe one of these nine types of artifacts. This number is increasing.

Conclusion

An estimate of the number of discrete knowledge artifacts described in articles indexed in PubMed since 1970 has an order of magnitude in the tens of thousands, not the hundreds or the millions. This number is rapidly increasing.

References

Towards Large-Scale Predictive Drug Safety: A Computational Framework for Inferring Drug Interactions Through Similarity-Based Link Prediction

Achille Fokoue, MS, Ping Zhang, PhD, Oktie Hassanzadeh, PhD, Mohammad Sadoghi, PhD
IBM T.J. Watson Research Center, Yorktown Heights, NY

Introduction

Drug-Drug Interactions (DDIs) may happen unexpectedly when more than one drugs are co-prescribed, causing serious side effects. Discovering and predicting DDIs will not only prevent life-threatening consequence in clinical practice, but also prompt safe drug co-prescriptions for better treatments. In this study, we propose a computational framework that takes in various sources of drug-related data and knowledge as input, and provides as output a list of potential DDIs along with an explanation for each DDI.

Methodology

We obtained a set of 12,104 drug pairs (among 1,197 drugs) already known to interact from DrugBank. By exploiting information originating from multiple linked data sources (e.g., DrugBank, UMLS, and CTD), we created 13 drug similarity measures to compare two drugs such as chemical structure similarity, pathway based similarity, and metabolizing enzyme based similarity. The resulting similarity metrics are used to build features for a large-scale and distributed logistic regression model to predict the existence of DDIs while efficiently coping with skewed distribution of DDIs and data incompleteness. The overview of our similarity-based DDI prediction approach is illustrated in Figure 1. In the experiment, we compared our method with a baseline that either assumes all drugs are known in advance during training or fails to account for skewed and rare events. Furthermore, we have developed a number of novel similarity measures that goes beyond the state-of-the-art approaches.

Results and Discussion

To fully evaluate our DDI prediction framework, we performed two types of experiments. Using the standard 10-fold cross-validation, our method achieved an average F-Score of 0.74 (vs. 0.65 for the baseline) and area under PR curve of 0.82 (vs. 0.78 for the baseline) for the newly developed drugs scenario (i.e., we held out all the DDIs associated with a fixed percentage of the drugs, rather than holding out DDIs directly). Additionally, we conducted a novel retrospective analytics to further demonstrate the effectiveness of our approach in DDI prediction. Our method correctly predicted up to 68% of the DDIs discovered between January 2011 to December 2014 by using only DDIs known before January 2011 as positive examples in training. This demonstrates our ability to discover valid, but yet unknown DDIs.

More details of our methodology and experimental results are presented in http://ibm.biz/adrtechreport.
Priorities for health-related use of personal data trails: Patients’ perspectives

Jacqueline Fontaine, MH1, Evette Ludman, PhD2, James Ralston, MD MPH2, Andrea L. Hartzler, PhD2
1University of Michigan, Ann Arbor, MI; 2Group Health Research Institute, Seattle, WA

Abstract

As the adoption of personal technologies grows, use of “personal data trails” from these tools promises to provide significant insight for research and care delivery. Yet use of patient-generated health data raises ethical, social, and legal concerns. Through focus groups, we investigated patients’ perceptions about sharing personal data trails for different health-related uses. Findings highlight concerns about personal autonomy, recipient trust, sharing costs and benefits, data quality, technological barriers, and limitations to law and policy.

Introduction

Widespread adoption of personal technologies, including smartphones, Internet, wearables, smartcards, and other ubiquitous devices, is accelerating the volume, type, and speed at which patient-generated health data spreads1. Personal data trails – digital footprints captured through our use of technology – carry insights into our characteristics and activities. This data also reflects our health, including what food we buy, how we exercise, and with whom and how much we openly interact. In this study we engaged patients to examine ethical, legal, and social issues surrounding the health-related use of personal data trails, including priorities of users who generate personal data (“generators”) when sharing their data trails with “recipients” in exchange for personalized services.

Methods

We conducted six 90-minute focus groups with 48 patients from Group Health Cooperative, a healthcare system in Washington State. Participants were adults who had actively used “MyGroupHealth”. Group Health’s Epic-based personal health record. Majority of participants were female (57.4%), White (77%), and well-educated (74.5% had a Bachelor’s degree or higher), with an average age of 59.8 (SD=16). Focus groups led participants through a series of scenarios that described how a fictitious patient “Susan” shared her personal data trail. Scenarios varied the personal data trail data types and recipients to explore differences in participant priorities in deciding when and with whom they share their data. Data types were expanded from Wood’s conceptualization of patient-generated health data and included lifestyle, online behavior, and retail data.2 Scenarios depicted this data shared purposefully with “primary recipients” and captured by third parties for secondary use. We coded transcripts using a codebook developed from initial qualitative themes, which were refined iteratively through collaborative coding among co-authors. Each transcript was independently coded by two researchers and compared for agreement. A third researcher coded transcripts to resolve disagreements. Any further coding discrepancies were resolved by discussion.

Results

Qualitative themes that emerged from the focus groups included general orientations towards sharing as well as costs and benefits of sharing and trade-offs when deciding to share. Foundational concerns that cut across these themes included the desire for personal autonomy and recipient trust. Other concerns were limitations of data quality, technology, and traditional laws and policies that do not necessarily safeguard health-related use of personal data trails outside the healthcare system. We are using findings to inform a conceptual framework to depict patients’ priorities for sharing across personal data trails across data types, recipients, and sharing pathways.

Conclusion

Patients provide a critical voice about the health-related use of their personal data trails. Themes about patients’ priorities depicted by our conceptual framework provides a foundation for future work that leverages patient-generated health data and informs thoughtful design and spread of personal health technologies.

References

Deriving the “Number Needed to Treat” from PubMed Structured Abstracts

Paul Fontelo, MD, MPH; Fang Liu, MS
National Library of Medicine, Bethesda, MD

Abstract
The Number Needed to Treat (NNT) indicates the number of patients needed to treat to prevent one unfavorable outcome. In some IMRAD-formatted abstracts, the Results section may contain adequate data of interventions to estimate the NNT. We developed a Web interface that automatically calculates the NNT after the table is manually populated with clinical trial results. A link is provided to the fulltext article if the abstract is insufficient.

Introduction
The Number Needed to Treatment (NNT) is an indicator of the effectiveness of a drug or intervention in preventing one adverse event in a patient population. Lower numbers are better. PubMed now has more than 16 million citations with abstracts and 3.1 million (19.8%) are structured abstracts. 406,000 are randomized controlled trials (RCT) and 51% of them are structured abstracts. Abstracts of some RCTs may provide enough information to calculate the NNT.

Methods
The PubMed for Handholds (PubMed4Hh) Web app algorithm was modified. Modifications involved two sets of vocabularies: the first set included identifying terms which indicate possible NNT terms, such as, “treated”, “control”, “intervention”, etc. The second set includes more accurate phrases, like, “treated group”, “control group”, “intervention group”, etc. RCT was set as the default publication type but other publication types are kept as options. Citations retrieved are scanned for the vocabularies in order and search terms are highlighted (left panel, Table 1). Citations with set 1 terms are marked with a star and those found by set 2 are marked with 2 stars, denoting that these sentences may contain data useful for completing the the NNT table. An NNT calculator in JavaScript is embedded in every abstract. Users manually complete the requisite data from the abstract on a 2x2 table. The NNT tool will calculate the Experimental Event Rate (EER), Control Event Rate (CER), Absolute Risk Reduction (ARR) and the NNT.

Result
The tool was tested on a set of 10 clinical queries. The tool performed as designed and citations correctly identified the terms in both sets (Table 1, right panel). Limitations include the requirement to manually complete the 2x2 table and that abstracts may not contain the data needed to complete the table. A link to the fulltext article is provided if the abstract does not include the data needed. The resulting values from the NNT calculator were the same as those from published manual methods.

Conclusion
An NNT calculator embedded in PubMed4Hh might be useful for some healthcare professionals to determine the NNT at the point of care. Development to automatically populate the table and use odds ratios are ongoing.
Factoring near-field chemical exposure into personalized medicine
Henry A. Gabb, PhD and Catherine Blake, PhD
University of Illinois at Urbana-Champaign, IL, USA

Introduction
Personalized medicine considers many factors (e.g., diet and genetics) affecting how a patient responds to treatment. However, the effect of long-term chemical exposure from consumer products is not typically taken into account. Though not acutely poisonous in normal usage, long-term exposure is potentially harmful and could exacerbate existing medical conditions and could affect how medication is metabolized\(^1\),\(^2\). For example, the fragrances and parabens in consumer products can exacerbate asthma. Roughly 80,000 chemicals are currently registered under the U.S. Toxic Substances Control Act. Most have not been subjected to toxicological risk assessment and fewer have been studied for potential drug interactions.

Materials and Methods
A database of 38,975 consumer products and their ingredients was compiled from online retailer sources\(^3\). Ingredient names were normalized using PubChem\(^4\) and the Unified Medical Language System\(^5\) in order to account for chemical synonymy and homonymy\(^6\). Product ingredients were compared to four authoritative lists of potentially harmful chemicals (Tox21, HSDB, EDCDB, and DODSON)\(^2\),\(^6\) to gauge their prevalence and distribution.

Results
Twenty-two of the 10,921 unique chemicals are shared in all four lists (Figure 1). Of these, 12 appear in the consumer products database: octinoxate, benzophenone, and benzophenone-3 (UV filters); octamethylcyclotetrasiloxane (cyclosiloxane); acetyl methyl tetramethyl tetralin (fragrance); diethyl phthalate and di-N-butyl phthalate (phthalates); triclosan and triclocarban (antimicrobials); and butyl, ethyl, and methyl paraben (parabens). The prevalence of these chemicals within five broad product categories indicates that the potential for exposure is high (Figure 2), though exposure duration and likely dosage must be taken into account before considering health outcomes. Parabens, suspected endocrine disruptors with estrogenic and allergenic potential, are in all categories; particularly skin care, cosmetics, and hair care products. Antimicrobials are also found in every category, though they are not as common as parabens because they are being phased-out of consumer products over similar health concerns. The paraben and antimicrobial compounds are associated with asthma severity due to immune sensitization\(^7\).

Conclusion
In addition to genomics, a patient’s product usage profile, or “productome,” could affect treatment outcomes. The list of potentially harmful chemicals is long but the results above suggest that exposure via consumer products could be used to prioritize the chemical-treatment combinations that should be further explored and considered in personalized medicine.

References

This research is made possible in part by a grant from the U.S. Institute of Museum and Library Services, Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12 Socio-Technical Data Analytics (SODA).
Developing a predictive model for discharge delay in the Post-Anesthesia Care Unit

Rodney A. Gabriel, MD, Jihoon Kim, MS, Lucila Ohno-Machado, MD PhD
University of California San Diego, La Jolla, CA

Abstract
We developed a predictive model using logistic regression to identify patients at risk for prolonged post-anesthesia care unit length of stay. The final model contained 7 predictor variables with an area under the ROC curve value of 0.765 (95% CI [0.741, 0.788]) on the training set and 0.764 (95% CI [0.729, 0.799]) on the validation set. This model may be beneficial from both a cost and patient satisfaction perspective.

Introduction
A predictive model that can identify patients who are at an increased risk for prolonged post-anesthesia care unit (PACU) stays could help optimize resource utilization and staffing needs. Although previous studies based on hospital stays identified some predictors, there is not a model that utilizes various patient demographic and preoperative comorbidities that may affect PACU length of stay in the ambulatory surgery setting.1-3

Methods
We collected 4,151 cases requiring care from an anesthesiologist in the ambulatory surgery PACU at a single institution. The binary outcome was the occurrence of prolonged PACU length of stay, defined as admission lasting more than one standard deviation above the mean recovery room time. A total of 20 predictor variables were obtained from electronic health records. The presence of a comorbidity was based on the existence of corresponding ICD9 codes extracted from each patient’s record. The dataset was split randomly into a training set (67%; n=2,782) and a validation set (33%; n=1,369). A logistic regression model was built on the training set using a combined method of forward selection and backward elimination based on the Akaike Information Criterion (AIC). The trained model was applied to the validation set. Model performance was evaluated with the area under the ROC Curve (AUC) for discrimination and the Hosmer Lemeshow (HL) test for goodness-of-fit.

Results
The final model had 7 predictor variables that included: primary anesthetic type, surgical subspecialty, obstructive sleep apnea, hypertension, congestive heart failure, history of postoperative nausea/vomiting, and scheduled case duration. The model had an AUC value of 0.765 (95% CI [0.741, 0.788]) on the training set and 0.764 (95% CI [0.729, 0.799]) on the validation set, with no significant differences in AUCs between them (p=0.980). The model had good calibration for the data in both the training and validation set as indicated by HL p-values of 0.52 and 0.14, respectively.

Conclusion
Several predictive models have been used in anesthesia1-3. This is the first model that utilizes patient factors that are identifiable in the preoperative setting as predictors for prolonged PACU length of stay specifically in the ambulatory surgery suite. The predictive model using logistic regression utilizes variables that may be predetermined prior to surgery (i.e., known comorbidities such as hypertension or sleep apnea). By developing a model using this type of data, future systems can automatically identify at-risk patients before surgery and aid in the allocation of staffing and resources in the surgical suite. This may prove to be beneficial from both a cost and patient satisfaction perspective and will be studied in future work.

References
The Use of a mHealth Decision Tree Support Program for Epinephrine Auto-injector (EAI) Administration Training of Adolescents

Joel Gallagher, MD¹, Robert Rivera, MS², Asriani Chiu, MD¹, Tanvir Roushan, BS³, Golam Ahsan, MS³, Cheng Wen, MPH⁴, Christina Eldredge, MD², and Shiekh Ahamed PhD³

1. Medical College of Wisconsin, Department of Asthma, Allergy, and Immunology, Milwaukee, WI, USA; 2. Medical College of Wisconsin, Department of Family and Community Medicine, Milwaukee, WI, USA; 3. Marquette University Ubicomp Laboratory, Milwaukee, WI, USA; 4. University of Southern California, Keck School of Medicine, Los Angeles, CA, USA

Introduction: Anaphylaxis is a potentially fatal allergic reaction which requires immediate treatment with an epinephrine auto-injector (EAI). Adolescents make up a disproportionately large percentage of anaphylaxis fatalities – 53% of the subjects in one study (1). Knowledge alone was unlikely to persuade adolescents to use EAI when needed (2), and only prior usage correlated with increased comfort with future EAI administration (3). Compounding this problem is the risk-taking nature of the adolescents (5). Encouraging EAI carriage without any other intervention is ineffective in ensuring that EAI are appropriately used (7). Current methods of teaching EAI administration utilize didactic sessions. However, it has been shown in various settings that skills wane as time since the session increases (9-13). There also exists a notable disparity in training between urban and rural settings (14). Therefore, novel interactive training modalities that are easily scalable to audience size and geographic setting are needed.

To address this gap in education, our mHealth decision tree support program focuses on training with the use of case-based scenarios and an interactive decision support mobile health application. The ultimate goal is to engage adolescents at a deeper level in decision making, therefore providing improved anaphylactic emergency preparedness. The mHealth decision tree support program creates a novel interactive virtual learning environment that allows for repetition using life-like scenarios. Repetition without focused responses has been shown to be ineffective in the successful retention of skills (15, 16). In contrast, our simulation concept results in continuous reinforcement of concepts and provides prompt feedback, allowing users to learn from their failures in a safe environment and nurturing the skills to successfully manage real life anaphylactic events. Current anaphylaxis training tools lack this level of interaction and decision support.

Methods: During the summer of 2016, our IRB approved trial will use a 3-arm, parallel, randomized group pre-intervention and post-intervention study design with a Control Group, Intervention A Group, and Intervention B group. All subjects will first undergo a simulation with a series of scenarios as measurement of their baseline performance. Then, we will provide each group with their respective training modalities. The Control Group will receive the standard of care only (an EAI demonstration via a clinician educator in conjunction with pamphlets describing anaphylaxis). The Intervention A Group will receive the standard of care plus training via the mHealth decision tree support program, while the Intervention B Group will receive the standard of care plus training via an application available to patients by a prominent EAI manufacturer. Each participant will be asked to complete a survey of anaphylaxis knowledge, self-efficacy, and health-related quality of life, before and after their respective interventions. Each participant will be introduced to their respective training modality and allowed a one week “run in” period to utilize their respective modalities. After one week, they will undergo another simulation with a series of comparable scenarios.

Conclusions: Current training modalities for teaching EAI use are insufficient and inadequately prepare food-allergic adolescents for anaphylactic episodes. We anticipate that our intervention will provide a viable and valuable alternative, increasing EAI carriage rates as well as increasing food-allergic adolescents’ health quality of life through improved self-efficacy and an increased feeling of safety. Results will be reported in our Fall 2016 presentation. In addition, our mHealth decision tree support program has the potential of reaching adolescents in a variety of geographic settings, therefore providing rural telemedicine support.
Comparison of Emergency Department Throughput Visualizations

Swaroop Gantela, MD1, Todd R. Johnson, PhD1, Nnaemeka G. Okafor, MD, MS1,2, David J. Robinson, MD1, Brent R. King, MD3, Amit M. Mehta, MD1, Nathan R. Hoot, MD, PhD1, Charles L. Maddow, MD1, Vickie Nguyen, MS1, Adriana Stanley, MS1, Amy Franklin, PhD1

1The University of Texas Health Science Center at Houston, Houston, TX; 2Memorial Hermann Health System, Houston, TX; 3Nemours/Alfred I. duPont Hospital for Children, Wilmington, DE

Introduction

Increasing importance is being placed on hospital throughput metrics1, creating an opportunity for situation awareness dashboards to play a supportive role. We discuss the use of visualizations of Emergency Department (ED) throughput data to provide cognitive support for real-time comprehension of the current state of the ED. Three visualizations were examined for their ability to assist in tasks ranging from global searches for bottlenecks to focused identification of individual patients whose metrics require attention.

Methods

The same dataset was used for all visualizations. The data represent a snapshot of all patients in a simulated ED and contain the length of time each patient has spent in their current stage of care. Three visualizations are presented: (1) the distribution of pending times for each stage of care, (2) bar chart for each stage, each bar a patient with the height being the pending time, and (3) a jump plot showing transitions across set checkpoints of care where each curve is a patient, the deviation from the midline is the difference from the median, and positive deviations are to the right.

Discussion

The distribution charts allow for comparison of absolute medians across stages, convey bottlenecks with increased density, give salience to outliers, but are hindered by overplotting. The bar charts allow for easy comparison of total count in each stage, allow for easy height comparison of patients within a stage, but are not as easy for comparing medians across stages. The jump plot is a new style of graph requiring some learning to interpret, but it allows for easy comparison across stages by setting the median of each stage on the common midline, emphasizes the position relative to the median, emphasizes outliers, and reduces overplotting.

Conclusion

No one chart served all purposes, but each chart had strengths in different contexts, and a combination of charts may provide utility in ED status assessment. The distributional chart aided absolute comparisons, the bar charts aided in count comparisons, and the jump plots aided in global views, highlighting outliers, and finding bottlenecks.

References

Documentation of Patient Strengths in Electronic Health Records

Grace Gao, RN-BC, DNP1, Madeleine Kerr, RN, PhD1, Ruth Lindquist, RN, PhD, ACNS-BC, FAAN1, Karen A. Monsen, RN, PhD, FAAN1
1School of Nursing, University of Minnesota, Minneapolis, Minnesota

Abstract

A strength-based model with a whole-person perspective aligns with person-centered care practice and documentation. A systematic review was conducted to determine the current state of patient strengths documentation in electronic health records (EHRs). Clinical information in EHRs is typically structured by problem-based diagnoses, however, there is emerging documentation of formalized strengths attributes using the Omaha System, which may promote a holistic approach to clinical practice and documentation using a person-centered ontology.

Purpose

Effective health management of individuals calls for a holistic, person-centered approach in clinical care and documentation. A whole-person representation captures not only patient problems but also patient strengths.1 Strengths are an integral part of well-being that involves physical, mental, and social dimensions.2 Studies suggest that patients want their strengths and inner resources to be utilized by care providers.3,4 To better understand and inform practice of person-centered care and documentation, a systematic review was conducted of the current state of patient strengths documentation in the EHRs.

Methodology

The data, information, knowledge, and wisdom (DIKW) informatics framework was employed to develop this systematic review.5 Three scientific databases were used to conduct a systematic or advanced search: CINAHL, Ovid Medline, and PubMed. The following search terms were used: strength*, problem*, whole-person, wellbeing, electronic health record*, personal health record*, EHR*, and PHR*. The inclusion criteria were to include literature related to utilization and documentation of patient strengths and/or problems in EHRs and/or personal health records (PHRs). The exclusion criteria were to exclude literature written in a language other than English or published before 2000. 384 articles were returned with different combinations of designated search terms. All articles were screened through review of titles, abstracts, or full texts. 20 articles were selected for this review.

Findings

The vast majority of articles focus on problem-based diagnoses and practice. However, one emerging trend appears to expand problem-oriented EHRs to include a person-centered ontology with the use of standardized terms to capture both needs-focused and strengths-based care and documentation. Early reports of strengths documentation were found using a standardized interface terminology and ontology, the Omaha System. Results of two studies demonstrated the feasibility of using the Omaha System for whole-person documentation to capture perception of both problems and strengths.1,6 Findings illuminate the merits of a strength-based approach in chronic illness management and well-being.

Conclusion

Clinical information in EHRs is typically structured by problem-based diagnoses, however, there is emerging documentation of formalized strengths attributes using the Omaha System, which may promote a holistic approach to clinical practice and documentation using a person-centered ontology.

References

An Examination of Provider Satisfaction with Telemedicine

Robert Garcia, MS1, Adelakun Olayele, PhD1, Wencui Han, PhD2
1DePaul University, Chicago, Illinois, United States of America; 2University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois, United States of America

Introduction

For many organizations medical providers serve as the gatekeepers to adoption and their satisfaction is critical to the success of telemedicine projects (1). While satisfaction remains a complex process understanding the factors that influence it can help influence the design and implementation of projects (2). The goals of this research are to identify common constructs used for examining provider satisfaction with telemedicine. These will be used for the creation of an instrument for measuring provider satisfaction with telemedicine.

Literature Review

Although patient satisfaction with telemedicine has been commonly studied there remains a need to study provider satisfaction (3). Factors that influence patient perspectives of telemedicine may not apply similarly to providers (4). Yet there remain methodological challenges in examining telemedicine satisfaction such as deficiencies in construct development and issues with the meaning of “satisfaction” (3).

Methodology

A survey of the literature was conducted on the United States National Center for Biotechnology Information’s PubMed database for the keywords “telemedicine satisfaction”. 167 studies that included empirical measures and used previously validated instruments were examined. 6 instruments with a total of 106 questions were identified that were used to measure provider satisfaction. A 10 step procedure using a grounded theory approach was used to identify common constructs from these questions (5, 6). The resulting constructs were then matched to similar concepts identified in the literature.

Results and Discussion

Few of the studies examined measured provider satisfaction using previously validated instruments. In some cases, provider satisfaction was measured using instruments previously validated for patient satisfaction or other types of systems. In total 16 categories were identified. These include: Overall satisfaction, treatment, workflow effectiveness, patient-provider relationship, time, cost, information availability, system quality, ease of use, usefulness, reliability, security, relative advantage, facilitating conditions, re-use and recommendation. While there is support for some of these constructs in the literature there are others that need further defining. Although several categories were identified, there may be other factors that contribute to provider satisfaction but were not evaluated in any of the instruments examined.

Conclusion and future work

Future work will expand on the number of instruments tested and attempt to identify additional constructs. These constructs will be further defined using methods described in the literature (5, 6). These constructs will be used to develop an instrument for measuring provider satisfaction that will then be validated and tested at a partner institution.

References

A Feasibility Study of Answering Clinical Questions Using askMEDLINE at the Point of Care

Kyungsook Gartrell, RN, PhD1, 2, Gwenyth R. Wallen, RN, PhD2, Caitlin W. Brennan, PhD, APRN2, Cheryl Fisher, EdD, RN-B, PhD2, Paul Fontelo, MD, MPH1
1National Library of Medicine, Bethesda, Maryland; 2National Institutes of Health/Clinical Center Nursing Department, Bethesda, Maryland

Introduction
The Internet plays an increasing role as a resource for answering questions in clinical practice. However, the availability of trusted resources and lack of time are known factors that hinder the use of technology at the point of care. Wireless mobile devices may mitigate barriers to access and delivery of reliable clinical information needed to support clinical decision-making among clinicians at the point of care.

PubMed for Handhelds (PubMed4Hh) is a mobile application from the National Library Medicine (NLM) for finding health information from biomedical literature. It features TBLs ("the bottom line" summaries), abstracts, and links to full text articles accessible anywhere the Internet is available. askMEDLINE, one of its features, allows the clinician to search using free-text queries of PubMed. PubMed4Hh is available for iOS and Android devices. The purpose of this study was to test the feasibility of answering clinical questions posed by clinicians using askMEDLINE during clinical rounds at the Clinical Center (CC) of the National Institutes of Health (NIH).

Methodology
This is a collaboration between the Lister Hill National Center for Biomedical Communications, NLM and the NIH CC Department of Nursing’s Research and Practice Development section. This observational study was carried out by a postdoctoral research fellow shadowing clinicians during rounds for one month in the following units: Oncology/Hematology, Oncology/Surgical, Pediatrics, Medical and Surgical Telemetry, and Intensive Care Unit. To simulate real-time conditions, clinical questions raised during rounds were searched using askMEDLINE using an iPad. Search results were evaluated by reading TBL summaries or the abstract, and the most relevant “answers” were selected. Search terms were highlighted in the TBL and abstract for convenience. The process was done independently and the results were not shared with clinicians, hence, not used in clinical decision-making. The clinical questions and search results were recorded electronically on a server at the NLM. No personally identifiable information on the clinicians, patients or CC units were captured. This study was exempted from Institutional Review Board review and approved by the Office of Human Subjects Research Protections of the NIH.

Results
Sixty-six questions were collected during the study period. The average number of questions per day was 3.7 (range: 1 to 8). The majority of questions (90.9%) collected were generated by physicians and 9.1% were from nurse practitioners. Physicians included residents, fellows, and attending physicians in various specialties. More than three-quarters (77.3%) of questions were from the Intensive Care Unit team. More than two-thirds of questions (65.2%) were answerable from the results obtained at the point of care. Almost 35% of questions were not answerable because of lack of time to read the search results, and/or the need to revise the questions when the initial search outcomes were not satisfactory. Reading search results and modifying search terms and resulted in the resolution of about 26% of the initially unanswerable questions. More than half of questions (60.6%) were related to information about medication or treatment, while 21% were questions regarding diagnosis, and 12% were specific to disease entities. The least frequently asked questions (6%) were related to information on patient data or information on instructions for the patient.

Conclusion
Our results show that real-time search and retrieval of reliable and potential useful information for clinical decision-making is feasible at the point of care. Since the information found was not used for clinical decision-making (i.e., the search was done independently by the researcher and not shared with the clinicians), the next phase is to make the search results available to clinicians, and assess its usefulness.
Using Electronic Health Records to Create Morbidity Based Epidemiological Profiles: Closing the Local Morbidity Surveillance Gap in Public Health

Becky Gawelek, MPH¹, Chris Kippes, MS¹, David C Kaelber, MD, PhD, MPH²,³
¹The Cuyahoga County Board of Health, Parma, Ohio, ²School of Medicine, Case Western Reserve University, Cleveland, Ohio, ³Center for Clinical Informatics Research and Education, The MetroHealth System, Cleveland, Ohio

Background
Chronic diseases, including: heart disease; cancer; stroke; diabetes; and arthritis, are the most common and costly of all health problems but they are the most preventable.¹ Population based morbidity data (along with associated risk factors) is one of the largest information gaps that confronts local public health (PH) departments. The purpose of the project is to determine the feasibility of utilizing electronic health records (EHR) for local PH surveillance.

Methods
Through a partnership between a large county health department and a large county hospital system in northeast Ohio, electronic health records (EHRs) (Epic Systems, Verona WI) were used as the primary data source. Inpatient and outpatient records from 1999 to 2014 were initially considered. Select demographic information were used in the generalizability assessments comparing the EHR based population within the county with the 2010 U. S. Census Bureau population statistics for the county. To conduct these assessments, we considered any patient that had an encounter between 2008-2012. Records were geocoded (ArcGIS 10.2.2, Redlands CA) to assign a zip code to the patient. Comparisons in proportions were made between EHR and Census populations with respect to 18 different standard age groups (0-4 years old, 5-9 years old, etc…), two gender groups (male, female), two race groups (white, black), and 51 zip codes. Statistically significant differences were assessed using chi-square analysis (IBM SPSS 23, Armonk NY) and a p-value of 0.001. Given the large sample size (over 300,000 patients), absolute differences of greater than 1.0% were considered to be of interest.

Results
Despite statistically significant differences (p < 0.001), absolute differences of greater than one percent were only found in nine of the possible 36 different age-gender groups (see Figure 1) and nine of the 36 possible age-race groups (data not presented). Zip code level comparisons and disease specific morbidity measures are pending.

Discussion
With multiple healthcare (HC) systems in our community, some baseline differences were anticipated. However, we were encouraged with the number of comparisons that had less than a one percent absolute difference. While optimistic, additional work remains to understand the full or limited value of using this type of data to close the PH morbidity surveillance gap.

Conclusion
EHR records can be a useful tool for local PH surveillance especially for morbid conditions. To help increase generalizability, attempts should be made to include all health system records through the use of health information exchanges. Furthermore, partnerships of this nature between local HC systems and public health can provide valuable experience for public health staff to gain proficiency with working with large datasets. Additionally, it serves as a resource for HC systems that need assistance with non-traditional analysis.

References
RDF-Based Method to Uncover Implicit Health Communication Episodes from Unstructured Healthcare Data

Pericles Giannaris, MS¹, Zainab Al-Taie, MS¹, Nattapon Thanintorn, MS², Ilker Ersoy PhD², Chi-Ren Shyu, PhD¹,³, Richard Hammer, MD², Dmitriy Shin, PhD¹,²,⁴

¹MU Informatics Institute, ²MU Department of Pathology and Anatomical Sciences, ³MU Electrical and Computer Engineering, ⁴MU Department of Computer Science, University of Missouri, Columbia, Missouri

Abstract

We present an informatics pipeline to discover healthcare communication episodes from unstructured healthcare reports. Our methodology is based on Resource Description Framework (RDF), ontological modeling, and description logic inference for uncovering and quantifying implicit communication episodes. We conclude that inferring these episodes helps to re-engineer healthcare communication structures important for exchanging amplified information.

Introduction

Healthcare communication includes transaction of specimen and medical records, oral interactions, etc. However, not every healthcare communication episode is formally recognized risking inefficient healthcare services, finding these episodes mitigates that risk. Here, with a proof-of-concept study, we propose a method to infer implicit communication episodes from free text pathology reports. The preliminary results show that our method reveals implicit communication both as information and artifact exchanges.

Method

Our informatics pipeline (Figure 1) includes the following components: (i) Ontology, (ii) Vocabulary, (iii) RDF Triplifier, (iv) RDF Store, and (iv) RDF Inferencer.

Ontology: The ontology provides classes and object properties describing the specific relations of terms in unstructured medical text. The classes are Communication Episode; Actor: an entity that serves as a sender or receiver of healthcare communication; and Object: any entity of healthcare communication representing Information or Artifact.

Vocabulary: The vocabulary serves as the basis of the ontology with a domain-specific set of selected terms found in surveyed unstructured pathology reports. These terms represent instances in the ontological classes.

The RDF Triplifier: The RDF triplifier parses with natural language grammatical and syntactical techniques unstructured pathology reports to terms that abide by the ontology and the vocabulary. These terms are semantically marked and encoded to RDF subject, predicate, and object (SPO) triples representing healthcare communication networks.

RDF Store: The RDF Store keeps triples that represent healthcare networks; each node is either an actor or an object linked by arcs symbolizing predicates of communication activity.

RDF Inferencer: The RDF inferrer checks the consistency of the RDF statements with a reasoning engine and infers implicit communication episodes from the RDF statements. For instance, given two RDF statements S₁ - P₁ - O₁ and O₁ - P₂ - S₂ the RDF inferrer collapses these statements to one and infers S₁ - P₃ - S₂. In general, the RDF inferrer uncovers implicit communication episodes by computing paths between actors in RDF networks.

Uncovered Communication Episode: It represents the output of the RDF pipeline: S₁ - P₃ - S₂

Results and Conclusion

We processed a sample of fifty unstructured pathology reports and our methodology uncovered five implicit healthcare communication episodes. For instance, one of the communication episodes that has been uncovered by the RDF pipeline is PathologyDepartment Send ExcisionalBiopsy, ExcisionalBiopsy ReviewedBy Dr.InConsultation; by collapsing the statements it infers that PathologyDepartment Send Dr.InConsultation. Therefore, we conclude that our RDF informatics pipeline infers healthcare communication that is not formally recognized and represented as acyclic communication networks.

References


1405
Patient Perceptions of Test Result Notification via the Patient Portal

Traber Davis Giardina, PhD, MSW¹, Jessica Baldwin, BA¹, Daniel T. Nystrom, MS, AEEP¹, Dean F. Sittig, PhD², Hardeep Singh, MD, MPH¹

¹Houston Veterans Affairs Center for Innovations in Quality, Effectiveness and Safety, Michael E. DeBakey VA Medical Center, Houston, TX and Section of Health Services Research, Department of Medicine, Baylor College of Medicine, Houston, TX; ²School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas

Introduction

A recent Department of Health and Human Services rule now gives patients direct access to their completed laboratory test results.¹ Additionally, Meaningful Use prioritizes patient online access to health information. While patient portals are now being implemented to provide patients access to their health information, little is known about patients’ needs when accessing their test results via the portal. Test result access is of high interest to patients and provides an opportunity to involve patients in preventing abnormal test results from being missed, a common patient safety concern. We conducted a mixed-methods study to explore patients’ experiences when accessing their test results via patient portals.

Methods

We recruited eligible participants (adults who viewed a test result in their portal) at four settings: an academically affiliated private general practice, a primary care community clinic, a network of private-practice physicians and an urban Veterans Affairs facility. We posted fliers at each site and a research coordinator approached patients in each clinic for potential recruitment. To boost sample size, we obtained a list of all portal users at one site and sent recruitment letters to a random sample of 450 patients (data collection is ongoing). Interviews were transcribed and then coded independently by the first and second author using content analysis. Descriptive statistics were used to summarize the survey data.

Results

Thus far, we have conducted five semi-structured interviews, three observations, eight structured interviews, and 33 surveys with patients. Our sample consists of 50% women and is 51.6% white, 28.1% black, and 12.5% Hispanic. Nearly two-thirds (65.6%) had one or more chronic illness and 57.8% were 50 years or older. In interviews and observations, use of the portal for viewing test results emerged as one of the dominant functionalities. While respondents valued access to their test results, many indicated that they needed additional contextual information to understand what the result means for their health and specific instructions for next steps for both normal and abnormal results. The majority of survey respondents were aware as to why the test was conducted (27.3% for routine visits and 57.6% for a specific health condition). To determine if the result was normal or abnormal, 40.0% looked at the range, 25.7% received the result from a clinician before they saw it in the portal, 6.1% saw a visual cue such as flag or highlight, and 6.1% were already familiar with the test. Although 81.8% indicated they understood the result upon viewing it in the portal, most sought additional clarification through online search (66.7%) and/or contacting their doctor by phone or secure message (51.9%). About one-third indicated feeling negative emotions, such as confusion or anxiety, after receiving the test result in the portal for both normal and abnormal results.

Conclusions

Our study, to our knowledge, provides one of the first mixed-method studies examining a new area of research, patients’ experiences of receiving test results through the patient portal. Providing patients access to their test results via portals should be accompanied by strategies to help patients interpret and manage results received. Additional work must devise strategies to leverage this technology to improve patient understanding of their overall health and next steps.

An Adaptive Conjoint Survey to Select Potential Features of a Technology Based Self-Management Intervention

Bryan S. Gibson, DPT, PhD1, Ana Sanchez Birkhead, PhD2, Elisa Amador, BSN2, Nancy Allen, APRN, PhD2

1University of Utah Department of Biomedical Informatics, Salt Lake City, UT; 2University of Utah College of Nursing, Salt Lake City,

Background: In order to maximize the effectiveness of technology based behavioral interventions, designers must ensure that the system’s features promote adoption and long-term engagement. Currently, the features included in most behavioral interventions are based on designers’ expertise or preferences, or qualitative input from potential end users. The goal of this project was to determine whether the adaptive conjoint survey method, which captures individuals’ relative preferences for features and attributes, could be used to inform design decisions in developing a technology-based self-management intervention.

Methods: To develop the list of potential intervention features, we held via a series of six focus groups with Hispanic community members with Type 2 Diabetes Mellitus (T2DM). In these meetings, attendees told their “diabetes stories.” The group then collaboratively identified common problems that people face in dealing with T2DM and discussed potential solutions. These solutions were then mapped to potential intervention features (including both well-established and novel functionalities).

We then developed an adaptive choice-based conjoint survey to determine which features were most desired and the modality by which people preferred they be delivered. The survey began with a “build your own” (BYO) section in which users chose the features they might be interested in (e.g., “something to keep track of my diet,” “something to monitor my activity levels,” etc.). Of the features that individuals chose, the adaptive survey then presented different possible configurations of the system using different modalities (e.g., reminders using text messages vs. email vs. phone calls). The survey was administered on a tablet computer in either English or Spanish.

In analyzing the data, we first calculated the mean importance for each potential intervention feature and then ran bootstrapped market simulations to ascertain the ideal (preference maximizing) configuration of features.

Results: 176 individuals initiated the survey and 135 provided complete data. Participants largely self-identified as Hispanic (98.5%), completed the survey in Spanish (94.7%), were mostly female (62.9%), middle aged (mean age 51 years, SD=13.3) and had limited access to healthcare: only 35.5% reported having health insurance. However, they had high ownership of TVs (97.7%), Smartphones (81.5%) and computers (63.7%).

The figure below depicts the mean importance for each potential intervention feature, notably TV ad replacement was the preferred amongst potential novel intervention features that our group is considering developing.

The market simulation, estimated that the ideal configuration of our potential system would address the preferences of 72.4% of potential end users. Most importantly, these simulations showed that whenever possible participants preferred functions via a smartphone.

Conclusion: Adaptive Choice Based Conjoint analysis provides a means to quantify end user preferences to drive system design decisions. In this project it helped to prioritize development of a novel intervention feature.
VIP: A Framework for Mining Clinical Concepts Using Knowledge Author Ontologies and Apache UIMA

Thomas Ginter, BS\textsuperscript{a,b}, Lalindra De Silva, MS\textsuperscript{a,b}, Olga V. Patterson, PhD\textsuperscript{a,b}, William Scuba, BS,\textsuperscript{b} 
Wendy W. Chapman, PhD\textsuperscript{a,b}, Scott L. DuVall, PhD\textsuperscript{a,b} 
\textsuperscript{a}VA Salt Lake City Health Care System; \textsuperscript{b}University of Utah, Salt Lake City, UT

Abstract
Mining domain-specific concepts from large collections of clinical text is an increasingly common task that facilitates many clinical studies. While existing Natural Language Processing (NLP) systems allow concept extraction, not many of them enable the full spectrum of domain-specific concept definition and concept extraction. To this end, we present a framework where clinical concept definition is handled through a sophisticated clinical ontology definition tool and the concept extraction is provided through an integrated Apache UIMA NLP pipeline.

Background
Extracting domain-specific concepts from large collections of clinical corpora is frequently needed for many clinical studies. Common approach is to define a set of domain-specific concepts, which is often carried out by domain experts, using a domain ontology creation tool. While many of these tools cover broad domains, it is often beneficial to work with more domain-specific tools that can exploit external clinical resources in the concept definition phase. Knowledge Author\textsuperscript{1} (KA) is one such tool that enables easy definition of domain-specific clinical concepts and allows users to augment their concept ontologies using existing resources. However, KA does not include the capability for a downstream NLP application to extract instances of these concepts from large corpora.

Methods
In this work, we present a framework named VIP, which facilitates mining of KA-defined clinical concepts in large text collections. The general architecture of our framework is presented in Figure 1. The input to our framework is a KA-authored clinical ontology in OWL format. KA allows concepts to be augmented with synonyms, abbreviations, CUI codes, etc. using the UMLS Metathesaurus. The VIP Decipher module extracts concept variables and relationships from the OWL representation. It then serializes these variables into a Leo service definition file that is used by VIP to create an NLP pipeline utilizing Leo framework based on Unstructured Information Management Architecture Asynchronous Scaleout (UIMA AS).\textsuperscript{2,3} The Leo service defines a series of UIMA annotators using the resulting service in UIMA AS for processing of the clinical data. A Leo client is executed from the command line to pull data for processing and provide reporting and output functions once the service completes data processing.

Figure 1: VIP Workflow

Knowledge Author also allows the users to qualify and quantify the clinical concepts and their associated contexts. For example, a user can define the clinical concept fever, attach it to UMLS concept C0015967 fever and create a complex variable, which includes attributes, such as experiencer [patient or other] and temporality [historical, present, etc.]. To enable extraction of the concept instances that satisfy these contextual attributes, we also implemented ContextAnnotator – a UIMA annotator based on the Context\textsuperscript{2} algorithm - to filter and retain only the desired instances.

Conclusion
In this work, we presented a framework for integrating the process of defining clinical concepts with the process of mining those concepts in large text collections. Our framework exploits the benefits of Apache UIMA for creating NLP pipelines that can process large text collections and we highlighted its capabilities in retaining relevant concepts.

References
Migration to Electronic Documentation – Benefits Seen at a Community Hospital

Derek Goodyear D.O.
East Liverpool City Hospital, East Liverpool, OH, USA

Introduction
The past decade has seen a huge increase in the use of electronic health records (EHRs) in the US. The goal is primarily to improve patient care and safety, but also to save hospitals money\(^1\). This may occur through reducing the need for transcription services and by improving communication to patient care teams. Our goal was to see if using our EHR instead of a transcription service would show any such benefits.

Methods
Over the last year, East Liverpool City Hospital, a 110 bed community hospital in Eastern Ohio, has undergone a transformation in their transcription process. Until June 2015, all reports (H&Ps, consultations, progress notes, discharge summaries, etc.) were recorded through a phone dictation service. Starting in June 2015, a transition was made to electronic notes within the Meditech EHR. This was initially rolled out to the hospitalist service followed by general surgery and then to other specialties.

Training led by Information Services and myself was provided to the physician staff to assist in this process. During this time, there were multiple sessions where staff physicians would document on mock patients and decide on what was important to include in the Review of Systems, Physical Exam and other sections. Over a 3 month trial period, progress, operative and post-op notes were migrated from transcription to electronic notes, followed by H&Ps and discharge summaries.

![Figure 1. Total Line Count Use in 2014 and 2015](image)

Results
Records were kept of how many total lines were being dictated and there was a significant drop after the change. The total cost to the hospital also decreased, saving over $30,000 in a period of 6 months. Because of this cost savings, a renegotiation was made with the transcription vendor. Now the electronic documentation is being rolled out to more providers and departments within the hospital. With the new documentation procedures, the primary service and all consulted specialists know exactly what each other are thinking right after they submit their notes. Time spent by physicians contacting each other for status updates was perceived to have decreased without compromising patient care, although this was not specifically recorded.

Conclusions
EHR documentation leads to immediate publication of records to facilitate physician communication and improve patient care; in comparison, the dictation system can take up to 72 hours, delaying important communication between physicians. We also found that electronic documentation can save hospitals a significant amount of money each year.

References
The Impact of Emergency Department Census on the Decision to Admit

Jillian K. Gorski, BS¹, Robert J. Batt, PhD, MBA², Erkin Otles, BS³, Manish N. Shah, MD, MPH³, Azita G. Hamedani, MD, MPH, MBA³, Brian W. Patterson, MD, MPH³
¹University of Wisconsin School of Medicine and Public Health, Madison, WI; ²University of Wisconsin-Madison, Wisconsin School of Business, Madison, WI; ³University of Wisconsin-Madison, BerbeeWalsh Department of Emergency Medicine, Madison, WI

Introduction

The Emergency Department (ED) serves as a “gatekeeper” for inpatient services. Hospital admissions rose 4% from 2003 to 2009, and EDs were responsible for almost all of the rise in admissions¹. It is known that ED physicians’ decision to admit or discharge patients has a significant impact on treatment costs². Furthermore, as ED utilization rises, the phenomenon of crowding - the operation of the ED beyond its resource capacity - can be observed. As an impact of the Affordable Care Act, increasing attention is being directed to system-level outcomes, while physician responses to crowding factors, such as whether admission decision are affected, remain less well characterized. The electronic health record (EHR) stores a wealth of operational data, and this information has not previously been used to examine this relationship.

Methods

We utilized a novel operational dataset built from event times recorded in the electronic health record to determine the effect of ED crowding on admissions decisions made by ED physicians. We performed a retrospective analysis using 49,487 adult patient encounters taken over an 18 month time period at a Level 1 trauma center. Waiting room census and physician load were calculated at the time of physician assignment. Univariate and multiple logistic regression models were created to assess the association between these census variables and the disposition decision, controlling for potential confounders.

Results

Of the encounters included in the analysis, 37% resulted in a hospital admission. Both census measures showed a statistically significant positive association with admission; the OR (odds ratio) per patient increase for waiting room census was 1.011, (95% CI: 1.001-1.020), and the OR for physician load census was 1.010, (95% CI: 1.002-1.019). Table 1, below, shows the output of the regression model at increments of patient census, illustrating the combined effect of waiting room census and physician load on likelihood of admission.

Table 1. Predicted likelihood of admission at given census values in the combined multiple logistic regression

<table>
<thead>
<tr>
<th># of patients</th>
<th>Waiting Room Census</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Physician Load</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>35.3%</td>
</tr>
<tr>
<td>8</td>
<td>36.6%</td>
</tr>
<tr>
<td>16</td>
<td>37.9%</td>
</tr>
</tbody>
</table>

Conclusion

Our results indicate that increased crowding may result in potentially avoidable ED admissions. This suggests that disposition decisions in the ED are influenced not only by objective measures of a patient’s disease state, but also by workflow related concerns. Further, we were able to study this relationship exclusively utilizing data retrospectively abstracted from the EHR.

References


1410
Electronic Messaging During Patient Care Coordination

Erika Green, MD¹, John T. Finnell, MD MSc FACEP¹, Emily C. Webber, MD FAAP²
Regenstrief Institute for Biomedical Informatics, Indianapolis, Indiana; ²Riley Hospital for Children at Indiana University Health, Indianapolis, Indiana

Introduction
Many communication tools are used in healthcare with a mix of benefits and limitations. Paging systems are neither efficient nor secure, EHR Messaging tools are not universally available, phone/callback cycles disrupt workflow. Text messaging has become a popular choice for fast communication in private and business settings, and is also increasingly utilized by medical staff during patient care.¹ But despite the variety of tools available, healthcare communication continues to experience a significant tug of war between efficiency and security. This study seeks to 1) describe the communication tools used currently by team members involved in patient care coordination and 2) find provider attitudes toward the efficiency and security of each.

Methods
Web-based, anonymous, IRB approved survey was distributed to physicians and advanced care practitioners, pharmacists, and case managers at a 19 hospital health system with multiple ambulatory centers.

Results
704 (31% of all invited) team members responded to our survey. When contacting medical staff, pagers are used predominantly by all groups. Phones are the most common tool to communicate with other team members. Staff working in an outpatient setting utilize EHR Messaging twelve times as often as the inpatient group (26.25% vs 2.20%, p=0.05), while the inpatient group prefers to use text messaging (15% vs 8%, p=0.05) and pagers (52% vs 21%, p=0.05). Staff in outpatient setting also finds EHR Messaging to be the most efficient and secure tool. Overall, twice as many respondents chose text messaging to be most efficient (202, 34%) compared to EHR Messaging (109, 18%) or phones (118, 20%), however, overwhelming majority perceived EHR Messaging (300, 50%) to be the most secure available tool. Because communication best practices are still under development at this institution, the results reflect employee’s unguided use and preferences.

Table 1. Communication tool utilization and attitudes on security and efficiency.

<table>
<thead>
<tr>
<th>Group</th>
<th>Tool Used To Reach Physicians</th>
<th>Tool Perceived</th>
<th>Most Secure</th>
<th>Most Efficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>To Reach Non- Physicians</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical team by role</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physicians</td>
<td>Pager (111, 40%)*</td>
<td>Phone (75, 29%)</td>
<td>EHR Messaging (134, 52%)*</td>
<td>Text Message (85, 33%)</td>
</tr>
<tr>
<td>Pharmacists</td>
<td>Pager (81, 60%)**</td>
<td>Phone (58, 43%)</td>
<td>EHR Messaging (55, 41%)</td>
<td>Text Message (59, 44%)</td>
</tr>
<tr>
<td>Case Managers</td>
<td>Pager (15, 58%)</td>
<td>Email (46, 33%)</td>
<td>EHR Messaging (98, 70%)**</td>
<td>EHR Messaging (48, 34%)</td>
</tr>
<tr>
<td>Venue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>Pager (141, 52%)*</td>
<td>Phone (117, 43%)*</td>
<td>EHR Messaging (101, 37%)</td>
<td>Text Message (130, 48%)*</td>
</tr>
<tr>
<td>Outpatient</td>
<td>EHR Messaging (63, 26%)</td>
<td>EHR Messaging (26%, p=0.05)</td>
<td>EHR Messaging (152, 63%)</td>
<td>EHR Messaging (80, 33%)</td>
</tr>
<tr>
<td>Age of respondent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 18-30</td>
<td>Pager (35, 56%)**</td>
<td>Phone (28, 45%)</td>
<td>EHR Messaging (26, 42%)</td>
<td>Text Message (24, 39%)</td>
</tr>
<tr>
<td>Age 31-50</td>
<td>Pager (129, 43%)*</td>
<td>Phone (98, 32%)</td>
<td>EHR Messaging (154, 52%)*</td>
<td>Text Message (108, 36%)</td>
</tr>
</tbody>
</table>

*Substantial vote (at least 2x the response rate of the next highest vote)
** Overwhelming vote (at least 3x the response rate of the next highest vote)

Conclusion
Most of the communication tools utilized by inpatient team during patient care today are perceived to be neither efficient, nor secure. There is a significant gap in the inpatient environment between efficiency and security that may be narrowed by evaluating the role of secure text messaging. Efficiency and utilization may also be improved by clarifying standards for best use for each communication method.

References

1411
Automatic Workflow Extraction Using Sequential Pattern Mining and Electronic Medical Record Usage Data

Tim A Green¹, Michael Phinney², Chi-Ren Shyu, PhD¹,²,³,⁴
¹Informatics Institute, ²Department of Computer Science, ³School of Medicine, ⁴Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO

Understanding sequences of events is paramount to developing a picture of how clinicians interact with Electronic Medical Records (EMR). Applications for EMR workflow include measuring productivity to help target training, and dynamic EMR experiences based on predicted actions. Common methods of assessing clinical workflow include video recording, direct observations, or directed user experience testing in a controlled environment. Computational methods can also be applied, such as sequential pattern mining [1], workflow mining [2], and graph-based mining [3]. To apply these methods, reliable sources of trace data that mirror clinician behavior within the EMR must exist. Often this trace data can be collected from debugging log files of computer systems. Benefits of computational workflow discovery include elimination of the observer effect, a potential for greater amounts of data for analysis, and the ability to include timestamped data from other ancillary systems in observed workflows.

The EMR at University of Missouri Health generates log data that is suitable for use in computational workflow discovery. We developed a system that extracts meaningful frequent sequential patterns from a combination of EMR log data and Admit, Discharge, and Transfer (ADT) data. One day’s worth of EMR log data from the middle of the month, and one months’ worth of matching ADT data for outpatient clinic visits were used, consisting of approximately 1.25 million records. EMR and ADT data were collated into 14,219 sequences that mirrored a physician’s interaction with the EMR. We looked for patterns indicating a predilection for use of the EMR before or after a patient visit to a clinic. ADT records were tagged with categorical timing information as it related to EMR interaction (Table 1). ADT time categories were 1) within 24 hours, 2) one to seven days, and 3) greater than seven days before or after EMR use. Sequential pattern mining running on Apache Spark [4] produced 14,990 frequent patterns, with a minimum support threshold of 0.20. There were many interesting frequent patterns identified that show the most commonly occurring EMR modules used, and in which order. We focused on three meaningful workflow patterns that included an element related to ADT (Table 1). We discovered that all frequent patterns over the support threshold included use of the EMR after a patient arrived at clinic. No frequent patterns emerged showing EMR use before clinic arrival, indicating that using the EMR as a preparatory tool may not be occurring frequently, or by physicians.

In this preliminary work, we have demonstrated that this system is capable of identifying commonly occurring patterns of EMR use. Future work will include generation of workflow graphs based on discovered sequences using sequential pattern mining, and use of sub graph mining techniques to look for more complex workflow patterns.

References

A Scalable Dataset Indexing Infrastructure for the
bioCADDIE Data Discovery System

Jeffrey S. Grethe, PhD1, Burak Ozyurt, PhD1, Hua Xu, PhD2, Xiaoling Chen, PhD2, Ruiling Liu, MS2, Anupama Gururaj, PhD2, Hyeon-eui Kim, RN, MPH, PhD1, Yueling Li, PhD1, Claudiu Farcas, PhD1, Alejandra Gonzalez-Beltran, PhD3, Philippe Rocca-Serra, PhD3, Ian Fore, DPhil4, Ronald Margolis, PhD4, George Alter, PhD5, Susanna-Assunta Sansone, PhD3, Lucila Ohno-Machado, MD, PhD1

1Univ. of California, San Diego, CA; 2Univ. of Texas Health Science Ctr. at Houston, TX; 3Univ. of Oxford, UK; 4National Institutes of Health, Bethesda, MD; 5Univ. of Michigan, Ann Arbor, MI

Abstract
The biomedical and health care Data Discovery Index Ecosystem (bioCADDIE) is an NIH funded Big Data to Knowledge (BD2K) project which is building a prototype (DataMed; http://datamed.org) for data discovery analogous to PubMed. A core component of DataMed is a scalable dataset indexing infrastructure that aggregates metadata from federated repositories and community aggregators.

Introduction
The development of bioCADDIE was started in 2015 with three primary goals: 1) Help users find sharable data; 2) Build a prototype (software system) for Data Discovery, called DataMed to reflect the analogy with PubMed; and 3) Engage the community and interoperate with other systems. The indexing infrastructure of DataMed maps disparate metadata from the diverse data sources into a unified specification provided by various working groups organized by bioCADDIE and related communities. This pipeline involves an automated component that provides controlled translation and curation of metadata using special tools such as a transformation language and JSON-Path.

Overview of the bioCADDIE Dataset Indexing Infrastructure
Since many indexing projects already exist and many repositories are already maintaining detailed metadata about the data sets they host, our goal was to leverage all this work and build a cross-repository, cross-indices index. This umbrella system was designed to interoperate with these indices and repositories, and not to replace them. The overall infrastructure (see Figure) consists of the following components:
• A data and metadata extraction system that is able to connect to various repositories and data aggregators. All metadata information is converted to JSON documents for each dataset being described and stored in MongoDB.
• A messaging infrastructure, utilizing Apache ActiveMQ, distributes dataset description documents from MongoDB, and depending on their status value, dispatches them to persistent point-to-point queues.
• A collection of multiple concurrent consumers retrieve documents from MongoDB, process them, update the job status and save them back. Consumers can be written using the STOMP protocol. Documents are transformed, to align with the bioCADDIE metadata model, and processed by a collection of modules that enhance the metadata records.
• Fully processed documents are then exported to an ElasticSearch endpoint that serves the dataset indices via RESTful services.

Current Status
The indexing pipeline for DataMed currently contains dataset description documents for 649,055 datasets from 23 different repositories that include 10 different data types. Work continues to incorporate new repositories to expand the content as well as on validating, testing and improving the metadata descriptions.

Acknowledgements
bioCADDIE is funded by NIH U24AI117966 and has many more team members and collaborators than would fit the author list for this abstract. They are listed in our web site at http://biocaddie.org.
Evaluating Google Translate for Facilitating Survey Translation from English to Mandarin

Jia-Wen Guo, PhD, RN
University of Utah College of Nursing, Salt Lake City, Utah, U.S.A.

Abstract
This study evaluated the translation quality of Google Translate in translating a survey from English to Mandarin. A 33 item survey was used as an example to evaluate the outcome from Google Translate. An acceptable Google Translate outcome can be used a starting point for researchers although the translation quality was not highly satisfied. Cultural difference and the region where the language is used contribute to challenges in the translate process.

Introduction
A great need exists to translate surveys into the language of the culture being studied. The task of translation is not simply to translate word for word from one language to another. The main challenges are to produce a translated surveys that is the linguistic and cultural equivalent of the original and to find experienced and well-qualified translators within limited project time to complete the repeated forward–backward translation procedure. Machine translation systems can be used to support the iterative and time-consuming translation process. Google Translate, a free online machine translation tool, has been tested in clinical and research for facilitating the translation work, but it was unclear how it can support in survey translation.

Purpose
The purpose was to evaluate the translation quality of Google Translate in translating a survey from English to Mandarin.

Methods
This was an observation study from both quantitative and qualitative perspectives. The Pain Care Quality Surveys (PainCQ), a valid survey for measuring the quality of pain care with 33 items published in U.S. English,1,2 was used as testing material. The quality of translation including intelligibility of translations, structural accuracy, and usefulness of the translation was quantitatively evaluated by two nursing researchers (one from Taiwan and one from Mainland China—both fluent in English and Mandarin). Regarding the qualitative evaluation, three nursing researchers provided the feedback of the use of Google Translate.

Results
The Google Translate outcome showed the average level, 3 out of 5, on intelligibility of translations, structural accuracy, and usefulness of the translation. The number of word per question, the sentence complexity, and readability of the sentence did not show a statistically significant influence on the translation quality of Google Translate. However, the Google Translate outcome still can be used a starting point for researchers although it does not have a satisfied result of the translation quality.

Discussion
The challenges of translating the instrument from one language to another are complicated. A sentence with a few or many words may have the same challenge to Google Translate and human translators. Cultural difference and the region where the language is used contribute to challenges in the translate process.

References
The Effect of Regional Health Information Exchange upon Ambulatory-Care Sensitive Hospitalizations and the Potential for Post-Discharge Care Coordination

David A. Haggstrom, MD, MAS;1,2,3 Susan Ofner, MS;4 Brian Dixon, PhD;1,3,5 Dustin French, PhD;6 Michael Weiner, MD, MPH;1,2,3 Laura Myers, PhD;1 Susan Perkins, PhD4

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

Introduction: If health information exchange (HIE) has enough reach into the population, HIE programs may improve coordination of care through the sharing of medical information across health care systems. The VA/DoD Virtual Lifetime Electronic Record (VLER) program is an HIE program administered in cooperation with community partners, including the Indiana Health Information Exchange (IHIE). By sharing data collected about hospital admissions across different health care systems, HIE has the potential to improve post-discharge care. By sharing data about outpatient care, HIE has the potential to reduce ambulatory-care sensitive hospitalizations.

Methods: The study population included all Veterans who were seen at least once in the year prior to HIE implementation at a tertiary VA medical center in Indianapolis. Among these patients, all VA and non-VA (IHIE) hospital admissions were measured for 1 year before and 1 year after implementation of the HIE program. Using this information, we first identified the data source (VA and/or non-VA) identifying each hospitalization, and then summarized admissions at both the hospital and patient level because a single patient could be hospitalized more than one time. We also compared the rate of hospitalization for ambulatory-care sensitive (ACS) conditions among those who were and were not enrolled in the VLER program, before and after HIE implementation, using a “difference-in-differences” zero-inflated Poisson model.

Results: More than a third of Veterans’ hospital admissions were identified at both the patient and hospital level using non-VA data (Table 1).

Table 1: Source of Data as assessed at Hospital and Patient level

<table>
<thead>
<tr>
<th>Source of Data</th>
<th>Hospital Level</th>
<th>Patient Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n admissions(%)</td>
<td>n patients (%)</td>
</tr>
<tr>
<td>VA alone</td>
<td>15497 (53.5%)</td>
<td>8479 (63.6%)</td>
</tr>
<tr>
<td>Non-VA alone</td>
<td>9533 (32.9%)</td>
<td>3947 (29.7%)</td>
</tr>
<tr>
<td>Both VA and non-VA</td>
<td>3952 (13.6%)</td>
<td>895 (6.7%)</td>
</tr>
</tbody>
</table>

Veterans with admissions detected using non-VA data had a higher mean and maximum number of admissions than Veterans with admissions detected using VA data alone (Table 2).

Table 2: Mean, Minimum, Median, and Maximum Number of Hospitalizations at Patient Level

<table>
<thead>
<tr>
<th>Source of Data</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA alone</td>
<td>8479</td>
<td>1.83</td>
<td>1.58</td>
<td>1</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>Non-VA alone</td>
<td>3947</td>
<td>2.41</td>
<td>2.73</td>
<td>1</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>Both VA and non-VA</td>
<td>894</td>
<td>4.41</td>
<td>3.13</td>
<td>2</td>
<td>3</td>
<td>34</td>
</tr>
</tbody>
</table>

The difference-in-differences in the rates of ACS hospitalization (per 100,000 population) was not statistically significant among patients who were enrolled in the VLER program (4,694 before vs. 4,633 after implementation) compared to those who were not enrolled in the VLER program (3,090 before vs. 3,184 after implementation).

Conclusions: The high proportion of Veterans’ hospital admissions that can be detected using data from a regional HIE suggests the potential for care coordination after discharge. Making use of shared electronic data, pro-active health care teams may be able to improve medication reconciliation, completion of follow-up tests, and other post-discharge coordination activities across health care systems. However, the impact of HIE use in primary care upon ambulatory-care sensitive conditions was not detectable within the first year of program implementation.
Exploratory Analysis of Marketing Vs. Non-Marketing Tweets on E-Cigarettes

Sifei Han, B.S. and Ramakanth Kavuluru, Ph.D.

1Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky
2Department of Computer Science, University of Kentucky

Introduction: The broad topic of electronic cigarette (e-cig) use has become a major fault line among clinical, behavioral, and policy researchers who work on tobacco products. There are arguments on either side given their reduced harm aspect can help addicted smokers quit smoking while the long term effects of e-cigs are not yet thoroughly understood. Furthermore, based on recent news releases from the CDC, there is an alarming increase of e-cig use by never smokers particularly middle and high school students who might be acquiring nicotine dependence. Twitter has emerged as one of the most popular microblogging platforms since its introduction in 2006 especially among teenagers. A major amount of chatter on e-cigs on Twitter surrounds their marketing by vendors making it generally difficult to analyze regular e-cig tweets (from the general public) that are not dominated by marketing noise. In this effort, we build a text classifier that identifies non-marketing tweets on e-cigs and employ topic modeling techniques to extract popular themes in non-marketing tweets from a corpus of over 500,000 e-cig tweets. Our main objective is to accurately separate marketing and non-marketing tweets and explore topics in the latter group.

Methods: A total of 540,579 tweets were collected during a seventh month period (4/15 – 10/15) using Twitter’s streaming API with a set of e-cig related keywords: electronic-cigarette, e-cig, e-cigarette, e-juice, and e-liquid including variants where hyphens are replaced with either spaces or empty strings. From this dataset, both the authors independently hand-labeled a random collection of 1000 e-cig tweets as marketing or non-marketing. The labels matched for 87.3% of the tweets with an inter-annotator agreement of \( \kappa = 0.726 \) indicating substantial agreement. The disagreements for the 127 instances where labels didn’t match were resolved in a face to face discussion to obtain a consolidated dataset of 1000 tweets. We observed that 48.6% (95% confidence interval of [45.51%, 51.70%]) are marketing tweets; that is, close to half of the tweets were marketing related. We pre-processed tweets by removing stop words and replacing user mentions and URLs with two separate special tokens. We built a logistic regression classifier with n-grams, tweet vector representations [1] with dimensionality of 100, and numbers of hashtags and URLs as features. We computed average performance measures based on 100 distinct 80%-20% train-test splits of the dataset. Our model achieved a mean F-score of 88.7% (95% confidence interval [88.37%–89.04%]) for the non-marketing class. We applied our model to all the tweets in the dataset and conducted topic modeling based [2] and key phrase based analyses on tweets identified as non-marketing by our model.

Results: Our method resulted in a total of 269,538 (51% of the full dataset) non-marketing tweets. We manually analyzed the top 20 topics (note that topics are usually represented as ranked sequences of tokens) resulting from our topic modeling experiments. As expected, none of the topics were related to marketing e-cigs although there was one topic about e-cig reviews, the closest to a marketing theme. We found one topic on safety issues in using e-cigs specifically about exploding cartridges. There were three topics related to health policy and regulation of e-cigs, one each on epidemiological and business/financial aspects of ecigs, and five topics on health effects of e-cig usage including scientific research conducted in the field. We were not able to clearly ascertain themes for five topics. A similar analysis for marketing tweets revealed topics mostly on deals, promo codes, free shipping, and particular e-smoking devices. Given the concern that adolescents might be at risk of getting addicted to nicotine, we searched both marketing and non-marketing tweets with words such as teens, teenagers, kids, youth, child, and children. We observed that 3% of non-marketing tweets contain these keywords while only 0.15% of marketing tweets have them. This is not surprising since most of the marketing is done implicitly to appeal to adolescents (flavors, packaging) but the seriousness of the issue has deservedly gained traction in the general populace.

References

HIE Expansion Improves Identification of Frequent Hospital Visitors

Xiao Han, MD1, Tina Y. Lowry, MS1, George T. Loo, DPH1, Elaine J. Rabin, MD1, Zachary M. Grinspan, MD2, Lisa M. Kern, MD, MPH2, Gilad J. Kuperman, MD, PhD3, Jason S. Shapiro, MD1.

1Icahn School of Medicine at Mount Sinai, New York, NY; 2Weill Cornell Medical College, New York, NY; 3New York-Presbyterian Hospital, New York, NY

Introduction

Health information exchange (HIE) provides benefits through primary clinical use at the bedside and increasingly through various forms of secondary use of HIE data. The importance of HIE is increasingly recognized, but we are still far from achieving a nationwide HIE. Our objective was to show that increasing the size of an HIE network allowed increased identification of frequent ED users, demonstrating that data from increased HIE size has a secondary use in quality measurement that more accurately reflects the way that many patients access healthcare across multiple, otherwise unaffiliated provider organizations within a region.

Methods

Healthix is a regional HIE that comprised 31 hospitals in the New York metropolitan area and Long Island at the time of this study. NYCLIX, a precursor HIE and subset of Healthix, comprised 10 hospitals. Using de-identified Healthix data from 03/01/2009 to 02/28/2014, we compared the detection rate of frequent ED users from three sets of data: hospital-specific data (treating each hospital as an independent dataset), NYCLIX-wide data, and Healthix-wide data. All ED visitors were divided into three mutually exclusive cohorts: 1) high frequency (HF) ED users, who had more than 4 visits in any given 30-day period, 2) medium frequency (MF) ED users, who had more than 4 visits in any given year, 3) infrequency (IF) ED users, who had less than 4 visits in any given year. The results were compared for statistical significant using Chi square and Kruskal-Wallis.

Results:

Healthix (31-site HIE) had 3,627,490 unique ED patients and 8,078,624 ED visits; NYCLIX (10-site HIE) had 1,686,904 unique ED patients and 3,855,747 ED visits. Of the Healthix patients, 1.45% (52,544 patients) were HF users and 6.20% (224,848 patients) were MF users. Of NYCLIX patients, 2.00% (33,672 patients) were HF users and 6.47% (109,124 patients) were MF users. We identified 10.1% more frequent ED users (142,796 versus 129,637) using NYCLIX, and 18.6% more frequent ED users (277,392 versus 233,888) using Healthix, compared to hospital-specific data. During the study period, using Healthix, we were able to identify 409 patients with more than 100 ED visits, and 44 patients with more than 300 ED visit. However, there were no patients with more than 300 ED visits identified with hospital-specific data. The median numbers of total visits from the Healthix dataset were as follows: IF 1-2 (median[IQR]), MF 7-9, and HF 9-16 (p<0.001 with Kruskal-Wallis test). The percentage of patients who visited more than 1 ED were as follows: IF 9.4%, MF 40.5%, and HF 44.8% (p<0.001 with Chi square test).

Conclusion:

Expanding HIEs improves identification of frequent ED users, especially very frequent users with hundreds of visits. We suspect that similar results would be found when other measures that may be affected by patient crossover visits are applied to HIE datasets. Analyses such as these, and more advanced services like notifications, that use HIE as a new, community-wide, real-time clinical dataset may allow provider organizations and the healthcare system to allocate resources such as care management more effectively. The goal of this study was to provide evidence to support expanding HIE to larger nodes including more patients and hospitals. Improving detection rate of frequent ED users is an example of how larger HIEs can identify problems that can otherwise be missed by hospital-specific data or smaller HIEs.
Overcoming interoperability challenges of community-based health interventions

Saira N. Haque, PhD, MHSA, Shellery Ebron, MSPH, Westleigh A. Quattrone, MA, Robert Bailey, BA, Barry Blumenfeld, MD, MS

1RTI International, Research Triangle Park, NC, U.S.A

Abstract

This reports on analysis of three Health Care Innovation Award (HCIA) grantees to determine the factors influencing health information exchange (HIE). Findings indicate that these community-based health interventions could facilitate HIE by understanding and anticipating the impact of technical implementation, organizational support, and governance. Organizations leveraging HIE should familiarize themselves with how federal investment in HIE infrastructure, and local pilot activities, have supported using information across establishments to support care coordination.

Introduction

Community-based health interventions use information exchange to support population health. However, disparate systems make aggregating and using data from multiple sources challenging. The purpose of this study was to learn how three community-based health innovations overcame data interoperability challenges to facilitate HIE. The Health Care Innovation Awards (HCIA) were granted by the Centers of Medicare and Medicaid Services (CMS) to organizations with innovations designed to enhance the quality of healthcare, improve health outcomes, and reduce the cost of care through the use of care coordination/patient navigation, health information technology (HIT), and the delivery of preventive or health promotion services.

Methods

We used information gathered during the course of a larger program evaluation including grantee-prepared progress reports, site visits, and telephone interviews, which were then coded and thematically analyzed in NVivo, a qualitative analysis program. We analyzed data from 3 HCIA awardees whose interventions focused on health information exchange.

Results

We found that HIE was influenced by technical factors, organizational support, and governance. Technical factors included file size, fields not being used in the same way, errors in exchange and infrastructure. Interoperability challenges included stakeholders at different organizations interpreting standards differently, populating data fields differently across systems, and varying familiarity with information exchange across organizations exchanging data. Organizations can facilitate health information exchange by sharing with a central coordinating organization, having regular inter-organizational meetings and regular review of error reports. In addition, community-based resources such as implementation guides and coordinated training may facilitate HIE.

Conclusion

Our findings suggest ways in which healthcare organizations can facilitate health information exchange within their communities to facilitate community-based health interventions. Organizations are increasingly leveraging information exchange for population health interventions. Federal investment in HIE infrastructure, EHR uptake through the Meaningful Use Incentive Program and local pilot activities have supported using information across organizations to support care coordination. This work can aid policymakers and providers in developing and using technical and operational infrastructure needed to implement information exchange innovations. Progress toward interoperability and successful population health projects that leverage it have the potential to improve care, improve health outcomes, and lower the cost of health care.

References

Visual Optimization of a Provider Dashboard for Patient Reported Outcomes in Surgical Spine Patients

Mary Beth Hasselquist MD1, Elizabeth Austin MPH1, Brett Fey1, Liz Kellogg MPH1, Cynthia LeRouge, PhD1, Andrea Hartzler PhD 2, David R Flum MD,MPH 1, Danielle C. Lavallee PharmD, PhD1
1University of Washington, Seattle, WA   2 Group Health Research Institute, Seattle WA

Introduction
With the rising use of patient-reported outcomes (PROs) in clinical practice, there is an increasing need to understand the data visualization needs of clinical teams to support their use of PRO data for both individual patient decision-making and broader population health applications. Previous research has indicated that end users of PRO data have different needs with regards to data visualization based on how the data will inform decisions and the PRO workflow implemented within their clinical settings. The CERTAIN Hub is a patient-reported outcomes web-based platform developed by investigators at the University of Washington designed to collect and report PRO data (e.g., pain, function) and risk factors (e.g., smoking, age) for patients undergoing cervical and lumbar spine fusion surgery. One feature of the system is a prototype dashboard that displays aggregate, de-identified PRO data for use by clinicians and staff. Heuristic evaluation where usability experts review user interfaces against accepted usability principles is one form of usability assessment. This study showcases improvements resulting from heuristic evaluation to be considered in PRO design guidelines.

Methods
We conducted a heuristic evaluation (as part of a human-centered design approach to dashboard development) of the CERTAIN Hub dashboard to identify opportunities to optimize the visual display of PRO data. This evaluation was led by a visual design expert utilizing heuristic principles based on visual science, graphical perception and cognitive theory. Recommendations from the heuristic review were then contextualized with feedback obtained from users during their initial implementation period of the dashboard. Users were exposed to both before and after visualizations (see Figure 1).

Results
The heuristic evaluation led to several recommendations to improve the visual display of the dashboard and the accessibility and interpretability of the dashboard’s data. Two graphs were displayed in new graphical formats (i.e. a line graph was replaced by a column graph) to more appropriately display the PRO data in the way that complemented how clinicians were applying the data in clinical practice. Additional revisions were made to several graphs to improve readability and facilitate interpretation of the data by clinical teams, including changes to the color scheme of the graphs and the addition of a Minimal Clinically Important Difference (MCID) reference line. Users expressed preference for the revised visualizations.

Conclusions
Usability principles from other visualization contexts hold in the PRO context. Integrating heuristic evaluation usability principles contextualized for PRO into the design of PRO systems yields better dashboard design.
Business Process Modeling: Capturing the Workflow of Medicine

Peter J. Haug, MD1,2; Herman Post, BS1; Joseph Bledsoe, MD1,2
1Intermountain Healthcare, Salt Lake City, Utah; 2University of Utah, Salt Lake City

Introduction
Modern businesses face a constant challenge due to continuing changes in the business environment and the increasingly complex efforts necessary to plan, document, and manage the workflows associated with these processes. Business Analysts are continually developing models for new workflow processes. These workflows involve participation from a variety of individuals, departments, and external partners. The challenge of managing these processes has led to a number of tools that can be used to design and implement optimum workflows.

The delivery of medical care in many ways parallels the variety and complexity of processes in the business world. Many diagnostic and therapeutic activities are defined by complex, multistep processes involving a variety of caregivers across a multitude of care settings. A core challenge for any medical enterprise is to understand and manage the workflows necessary to deliver care.

In this poster presentation we will present and describe an application in which we have taken a tool originally designed to support business process modeling and adapted it to the medical setting. We have implemented a system based on an OMG standard, Business Process Modeling Notation, version 2.0 (BPMN 2.0), in four local emergency departments. The goal of the system is to assist in the diagnostic workup of Pulmonary Embolism.

The software environment described provides an infrastructure in which to both author and deliver medical protocols. During the course of any protocol, BPMN 2.0 provides a number of ways to execute decision logic including “gateways” which implement various types of flow-control logic, “script tasks” which allow insertion of small pieces of code into the workflow, and “business rule tasks” with which a workflow author can invoke the services of a rules engine to implement more complex clinical logic. If additional inferencing tools are required, they can be configured as “custom service tasks”, added to the palette, and made available to the workflow author.

Experience
An automated workflow for diagnosing Pulmonary Embolism (PE) is available on computer desktops in four emergency departments within Intermountain Healthcare’s Urban Central Region. A reminder to use this tool is triggered upon receipt of an order for a test associated with PE (D- Dimer, CT Pulmonary Angiogram, Ventilation-Perfusion scan). The workflow begins by guiding the user through a series of algorithms designed to estimate the likelihood of PE (based on the Pulmonary Embolism Rule-Out Criteria2 and the Revised Geneva Score3). It combines BPMN Gateways, Service Tasks (for data access), Script Tasks and User Tasks with some custom services to navigate the logic necessary to choose the most appropriate sequence of diagnostic tests. It concludes when sufficient information is collected to confirm or deny the presence of PE.

The pulmonary embolism protocol was implemented in four emergency departments within Intermountain Healthcare. Between June 2014 and February 2016, the prototype system underwent testing. The protocol was completed for 1084 patients; it was exited before completion in 620 cases. In many of the latter, the clinician reviewed suggestions before exiting. Of 2471 total suggestions displayed to the user (suggest D-dimer lab test, suggest imaging examination in general, and a suggestion for a specific imaging examination), 2353 were accepted and 118 were not accepted.

Discussion
Available technologies developed to design and implement workflow processes for the business community can be used to build a powerful environment in which to develop complex, stateful, clinical protocols. We are engaged in an effort to adapt these technologies for use in the medical environment. Initial experience has been gained through the development of a Pulmonary Embolism diagnostic workflow. Our experience suggests that these tools will ease the design, implementation, and long-term maintenance of electronic protocol that are capable of providing robust decision support as a part of an interactive computerized workflow.

References
Roadmap for Integrating an EMG Machine with the EMR

David Haustein, MD\textsuperscript{1,2} and Brenton Bohlig, MD\textsuperscript{2}
\textsuperscript{1}Robley Rex VA Medical Center, Louisville, KY; \textsuperscript{2}University of Louisville, Louisville, KY

Abstract

Electrodiagnostic testing, also known as a nerve conduction study or electromyogram (EMG), is a medical procedure that can help diagnose nerve or muscle diseases by using a stand-alone laptop or desktop computer equipped with EMG software, a preamplifier and a stimulator. Integrating this device with the electronic medical record (EMR) can be an expensive and daunting task for physicians; this poster will explain the cost and steps to upload the report directly into the VA’s EMR.

Project

Electrodiagnostic testing (EDX) or an electromyogram (EMG) is a medical diagnostic test that can document abnormalities after a nerve or muscle injury and is frequently used to confirm common conditions such as carpal tunnel syndrome, a pinched nerve in the neck or back or nerve damage in the hands or feet. EMG machines are stand-alone laptops or desktops with a preamplifier and stimulator and are infrequently connected to the electronic medical record (EMR). EMG machines that integrate with an EMR are reported to improve workflow efficiency, reduce clerical errors and provide faster communication of reports to referring providers. However, this integration requires extra software and hardware expenses and requires significant coordination with health information technology professionals. When the manufacturer no longer supported the older EMG machines at the Louisville VA, we began to research systems and the process for integrating EMG machines with the VA’s EMR, VistA. We first needed to choose an EMG machine that was an approved device through VA Clinical Procedures and also needed confirmation of interoperability with the local VistA Coordinator. For a fully integrated system with both importing of patient demographic data into the EMG machine and exporting of reports to VistA, Health Level 7 (HL7) software was required at an additional cost of $5,000 to $10,000. Finally, it was necessary to work with IT and biomedical engineering to obtain a server on-site to host the software and data. Many EMG manufacturers offer reader software which allows a provider to manipulate the EMG document from a remote workstation for about $1000 per workstation; this was important for our facility where we have multiple attendings and residents working in the EMG lab daily. Going live with this new process first required several weeks to get comfortable with the new EMG machines and then many administrative/health IT meetings to coordinate the vendor’s remote access to our on-site server for set up and maintenance. Our experience with this transition will serve as an informative roadmap for other clinicians of the prices and personnel involved in integrating EMG machines and their EMR.
Design Considerations for Patient Portals: Needs and Preferences from Elderly Adopters

Yuqi He, MLIS
University of Wisconsin-Madison, Madison, WI

Abstract
Patient portals adoption creates great opportunities and challenges for the growing aging population; focus group interviews with people age over 65 were conducted to understand their needs and preferences for patient portals and design considerations were presented based on the initial findings.

Introduction
In response to Stage 2 of the Meaningful Use under the HITECH Act, health care providers across the U.S. are required to implement patient portals and demonstrate a certain amount of portal usage to qualify for financial incentives. In addition to the law, patient portals are considered a promising tool to manage personal health information and improve patient engagement. A systematic review reported that portal use is associated with improvements in disease awareness, medication adherence, patient-doctor communication, and self-management of disease. Patient portals are believed to be beneficial for older adults because the increased occurrence of chronic conditions demands disease self-management and health information management. However, these portal benefits will only be realized if the portals are well designed and are used effectively by older patients.

We lack knowledge on how to make patient portals accessible for older adults and how to design them taking into account the needs, preferences, and capabilities of older adults (e.g., cognitive, perceptual, and motor abilities). To help design better information technologies that promote personal health information management and patient engagement, this study, now in progress, explores the range of older adults’ needs and preferences for patient portals regardless of their current adoption status. Health institutions that implement patient portals and IT designers may benefit from design considerations.

Methods
Focus group interviews are employed as the main data collection methods because ideas tend to emerge from groups and this method elicits a wealth of descriptive data. Groups are separated into elderly portal users and non-users to allow for more detailed discussion. Participants are primarily recruited from senior centers and public libraries. In current phases, two focus groups of portal users have been conducted. Eleven subjects (3 males, 8 females) aged 65 or older participated. Each focus group session lasted an average of 90 minutes. Participants’ demographics, computer proficiency, and portal use were assessed by three questionnaires. The interviews were recorded and transcribed verbatim. Notes were also taken at the time of the focus group to capture verbal content and nonverbal behaviors. Transcribed interviews and the interview notes were analyzed using Nvivo 10. The basic principles of constructivist grounded theory guided the data analysis.

Early Results
This presentation reports preliminary results on elderly portal adopters’ needs and preferences. Almost all participants felt positive towards the portal they were using, which were primarily versions of Epic MyChart. They mainly used the patient portals to send messages to the doctor’s office, make and verify doctor’s appointments, check lab results, and gain proxy access to a family member’s portal. Four themes emerged: (1) older adults would like various formats of training on how to use the portal, (2) they would like push notifications when new information comes to the portal, (3) additional resources are needed for understanding unfamiliar medical terminology, (4) managing multiple portal accounts from different providers makes portal experience frustrating.

Implications
The findings point to design considerations for older adults: (1) provide various formats of training (e.g., small group training workshops, video training, and written manual) to cater to older adults’ diverse computer experience and learning preferences; (2) provide options for notification methods and what kind of information should be pushed; (3) provide helpful and relevant information resources. Older patients desire an integrated health information system that provides a comprehensive approach to care. This forces us to think ahead in terms of not only meet the short-term meaningful use goal but also create long-term strategic plans to enable integrated IT to span the entire care continuum.

References
Assessing the Impact of Mobile-based Intervention on Health Literacy among Pregnant Women in Urban India

Aparna Hegde, MD1, Riddhi P. Doshi, MBBS, MPH2
1ARMMAN, Mumbai, Maharashtra, India; 2University of Connecticut Health Center, Farmington, CT, USA

Introduction

Every year, more than a fifth of the half million global maternal deaths occur in India.1 In 2012-13, almost half of the pregnant women in Mumbai received complete antenatal care (ANC) including three ANC visits, one tetanus toxoid injection and 100 iron folic acid supplements.2 Over 2/3rd of the pregnant women (15-49 years old) in Mumbai suffer from anemia. Only 69.2% of the children under 2 years of age received complete immunizations as recommended by the World Health Organization.3

Methods

This quasi-experimental study involved a single group, pre-post evaluation design. Of the 400 women recruited at Lokmanya Tilak Municipal General Hospital, 247 responded to the follow-up survey at 9 months. The intervention through Project mMitra was a free mobile voice call service that sends timed and targeted preventive care information messages directly to the mobile phones of enrolled women throughout pregnancy and infancy in their preferred language and time. The messages were developed by expert group of physicians from FOGSI (Federation of Obstetrics and Gynaecological societies in India), IAP (Indian Association of Pediatrics), and NNF (National Neonatology Forum) and modified based on focus groups studies. For this study, the voice messages included information regarding anemia and nutrition during pregnancy, breastfeedi ng, and immunization. The baseline and follow-up surveys were conducted by research staff to assess women’s awareness and understanding of health care during pregnancy and infancy. Chi square tests were performed to compare the changes in health literacy.

Results

At baseline, only half of the women had their own mobile phones, whereas over a third used a family member’s phone. A majority of women were skilled in receiving calls but only 56% knew how to send a text message. The proportion of women aware that iron folic acid supplementation is to be taken for ideally 100 days (p<0.01) and minimum 90 days (<0.01) improved significantly after the intervention. Awareness regarding ideal duration for exclusive breastfeeding (p<0.01) and importance of colostrum (p<0.01) also improved drastically. Almost 97% of women were aware about the importance of immunization during pregnancy upon follow-up compared to the 61% at baseline. About 93% respondents were aware about immunization schedule during infancy during the follow-up compared to 71% at baseline. Over 95% of the respondents were satisfied with the frequency, timeliness and relevance of the voice messages.

Discussion and Conclusion

This study demonstrates the feasibility and value of delivering mhealth messages to pregnant women in order to improve health literacy regarding preventive care during pregnancy and early childhood in urban India. One of the major limitations for this study is the potential for respondent bias. We have attempted to overcome this shortcoming by adopting the randomized controlled trial design in future studies.

References

Efficacy and types of health information technology used in diabetes education for medically underserved adults: A systematic review and meta-analysis

Elizabeth M. Heitkemper, MSN, RN¹, Lena Mamykina, PhD², Jasmine Travers, MSN, RN¹, Arlene Smaldone, PhD, RN¹

¹Columbia University School of Nursing, New York, NY; ²Columbia University School of Biomedical Informatics, New York, NY

Abstract

A systematic review (SR) and meta-analysis (MA) to examine the effect of diabetes education health information technology (HIT) interventions on health-related outcomes in the medically underserved was completed. Five databases were searched and 13 studies met criteria for SR inclusion and 11 for the MA. Results show that HIT is effective in improving blood sugar at 6 months with dissipating but significant effects in 12 months in underserved adults with type 2 diabetes.

Introduction

Diabetes impacts 29 million individuals in the United States and is the 7th leading cause of death. Health disparities illustrated by higher disease prevalence, poorer glycemic control and more diabetes-related complications are seen in racial and ethnic minorities, low-income and rural populations, known together as the medically underserved (MU). Successful management of diabetes entails lifestyle modifications and long-term behavior change, which are taught via individual or group-based diabetes self-management education (DSME). However, for MU populations, many do not receive DSME due to the considerable barriers that exist such as language and literacy, health beliefs and cultural considerations, and access. One timely solution to this problem is the use of health information technologies (HIT) since studies show increased adoption of technologies by groups that previously had little to no access makes HIT an increasingly viable option for the MU to receive DSME.

Methods

Following an a priori protocol and the Preferred Reporting Items for Systematic Review and Meta-Analysis guidelines, five databases were searched. Any duplicate publications using the same sample, opinion pieces, literature reviews, other systematic reviews and meta-analyses, or articles published in non-peer reviewed journals were excluded. Studies reporting either A1c pre and post intervention or its change at ≥6 months were eligible for inclusion in the MA using random effects models.

Results

Of 318 articles identified, 13 randomized trials met criteria for SR and 11 for the MA and represent data from 3,257 adults (mean age 55 years; 66% female; 74% racial/ethnic minorities) with diabetes. Interventions varied by HIT type: computer software with no Internet (n=2), cellular and/or automated telephone technology (n=4), Internet based (n=4), and telehealth/telemedicine (n=3). Most studies (n=10) reflected an unclear risk of bias overall. Pooled decreases in A1c were found at 6 months (-0.38 [95% CI: -0.50 and -0.26]; I²=35.1%, Q=5.0) with diminishing effect at 12 months (-0.24 [95% CI: -0.39 and -0.09]; I²=42.4%, Q=10.4). Heterogeneity was not greater than expected by chance alone.

Conclusion

The results from our study show that HIT DSME is effective in improving glycemic control at 6 months with dissipating but significant effects in 12 in underserved adults with type 2 diabetes. It will be important for developers to find novel ways to improve the technologies that this population has access to, such as smartphone applications, since our findings show that cellular or automated telephone technologies were not as successful as other technology types.
TransformDB: A Data Management System for Multimodal Imaging Data

Karl G. Helmer, Ph.D.\textsuperscript{1,2}, Qi Wang, M.D.\textsuperscript{1}, Graham C. Warner, B.S.\textsuperscript{1}

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

TransformDB was created to serve the Transdiagnostic Repair of Affective Networks by Systematic, Function-Oriented, Real-time Modeling and Deep Brain Stimulation (TRANSFORM-DBS) project (https://transformdbs.partners.org), which will generate a large, heterogeneous set of high-value neuroscience data. These data include brain electrophysiology recordings from humans and animals performing a variety of cognitive tasks and brain imaging (MRI and MEG) scans during similar tasks. These goals is to not just make the data available to the public, but to provide a semantically aware dataset that will enable discovery and meaningful reuse.

TransformDB is a registration-based system; both subject demographic as well as experimental data sets are registered into the system, either automatically or by users. The subject’s demographic data is entered upon initial contact and the subject is given a universal user ID (UUID). Data is registered at the level of a “run”, which is defined as an acquisition period of a single type (though this can involve multiple data files). Data with personal health information (PHI) is de-identified before transfer or storage and no PHI is stored in the system at any time. A compressed archive file of the data from each run is created and is transferred from a public-facing Globus (http://www.globus.org) shared endpoint to the file system where the metadata is extracted. The archive is then stored in the file system with a schema that reflects the experimental organization (project, subject, acquisition modality type, session, run). Currently, this process is performed automatically for MRI and MEG data, but manually for electrophysiology data due to the paucity of identifying metadata in those data formats. A comprehensive data and database integrity code library has been created to check for data and database integrity.

TransformDB is run on a CentOS (ver. 6.4) server that consists of a website front end and a CouchDB database (http://couchdb.apache.org/) and Flask server in the back end. The website is based on a Drupal framework. The Flask server is used to run Python scripts to perform data anonymization upon registration, data movement within the infrastructure, and the extraction and storage of metadata. The file metadata is both extracted from the file header, as well as entered by the user at the data registration step in the case of important information not in the file header, such as electrode type and manufacturer. The metadata records are fully query-able in both basic (e.g., subject demographic information and acquisition modality) as well as advanced (specific fields in the file header). Unregistered users are able to search the data collection; however, limited fields are returned. With registration, a comprehensive view of the data is available through the advanced search capability.

The data sharing infrastructure supporting the TRANSFORM-DBS project is based on a modern design meant to fully support the discovery and sharing of heterogeneous raw and processed data. At its core is the NoSQL document store, CouchDB\textsuperscript{2} that stores Javascript Object Notation (JSON) documents. These documents are ideal to handle the evolving nature of the data and the variety of file metadata in this project. JSON documents that can be easily updated and queried allowing for maximum flexibility as new data types are added, while minimizing query time.

Reporting tools are available for both internal and external use. At-a-Glance pages are available to projects noticing global views of subject data, displaying which scans have been completed as well as the tasks performed. Comprehensive reporting pages are also available so that users can determine the number of subjects with a given criteria set that exist within the data. Users can search for data, and create Shared Searches, named collections of data that can be stored for personal reuse or for transmission to others. By using a Shared Search, users do not have to physically transfer the data to others, with the Shared Search name, recipients can download the data themselves. These collections can also be used to create fixed collections of data that are independent of the state of the database and the addition of new data.

The terms used to describe the data are chosen based on our ongoing work with the INCF Program on Standards for Data Sharing’s Neuroimaging Data Model (NIDM) that build on previous ontological work in neuroimaging. The goal is to use standardized terms so that this data collection can be more easily queried with other collections using the NIDM data model thereby enhancing the data’s visibility. Each term is also defined to ensure comprehension of the data and to promote its reuse.
Evaluation of Clinical Score Calculator Programmability within the Electronic Health Record

Vitaly Herasevich, MD, PhD, Mikhail A. Dziadzko, MD, PhD, Christopher A. Aakre, MD
Mayo Clinic, Rochester, MN

Abstract
Clinical scores are frequently used in clinical practice, however many are cumbersome and require manual entry of multiple variables into web-based score calculators. We performed a preliminary programmability assessment of 171 commonly used, evidence-based clinical score calculators to guide score automation efforts within the electronic health record.

Introduction
Although a few clinical scores have been integrated into electronic health record (EHR) platforms, most score calculation is still performed manually by clinicians - assisted by smartphone or web-based applications. These other scores are used frequently in clinical practice for decision support, risk prediction, and disease severity assessment. Automation of these clinical score calculators within the EHR has the potential to save clinician time and to improve the quality of clinical care. However, the ability to automate each score depends on its programmability; programmability has not been assessed for most clinical scores. Filling this knowledge gap would help to focus research on future EHR-based score calculator automation.

Methods
Variables from 171 clinical score calculators identified from the medical literature in a prior study(1) were tabulated. Each variable was classified as either “subjective” or “objective”. Objective variables included laboratory values, demographic values, test results, vital signs, and regularly charted standardized data. Subjective variables included elements of clinical history, diagnoses, examination findings, and clinical judgments. Subjective variables with established advanced retrieval techniques, such as natural language processing, were also identified. Programmability was defined as the proportion of objective variables.

Results
1343 variables (618 unique) were tabulated from 171 clinical score calculators. 270/618 (43.7%) unique variables represented objective data present in the EHR. An additional 114/618 (18.4%) of variables are potentially extractable with advanced retrieval techniques. 23/171 (13.5%) clinical scores were completely programmable and 73/171 (42.7%) of clinical scores are composed of 75% objective variables or greater. 31/171 (18.1%) contained fewer than 25% objective variables. When advanced retrieval techniques are applied, 50/171 (29.2%) could be completely programmable and the number of scores with 75% or greater programmability increased to 108/171 (63.2%). A summary of the programmability of the 171 scores can be seen in Figure 1.

Conclusion
More than 60% of commonly used clinical score calculators are highly programmable (> 75%) for automation in the EHR using objective data and advanced variable extraction techniques. Further research into automation of complex, highly used, highly programmable scores within the EHR could save significant clinician time.

References
Inconsistency and Variation in Concordance of Social History Documentation between Nursing and Provider Clinical Notes

Jamie S. Hirsch, MD,1,2 Kevin Bock, MD,1,2 John Chelico, MD,1
Petros Karamanlakis, MD,1 Deborah Mensch, MD,1,2 Michael Oppenheim, MD1,2
1Northwell Health, Lake Success, New York;
2Hofstra Northwell School of Medicine, Hempstead, New York

Introduction
Electronic health records (EHRs) with clinical documentation are widely implemented across many health systems. As a primary and major source of rich information, clinical notes provide knowledge of a patient’s social history and context, with implications for patient care as well as quality improvement and regulatory reporting. To aid in documentation consistency and generation of reusable data, notes often contain structured, discrete data entry forms, with many elements identical in documents entered by several members of a patient’s health care delivery team. To what degree this information is reliable and concordant across this care team, however, is unknown.

Methods
In order to improve documentation of crucial variables the EHR at Northwell Health affords the ability to document similar observations among members of a health care team. An additional option allows charted observations to carry over from previous notes. We examined the social history portion of provider and nursing admission notes across 8 hospitals at Northwell Health for adult admissions in 2015. Focusing directly on elements with discrete data entry, we examined the proportion of observations completed by providers and nurses in their respective admission documents, and when both completed a similar observation, the degree of concordance in the valued observation.

Results
In 2015, there were 95597 non-elective, adult admissions with a provider and nursing admission note, with at least one containing documentation of social history. The median number of social history observations documented was 5 for providers and 4 for nurses (IQR, 4 and 5, and max 46 and 45, respectively). While some observations are unique to only one document type, most (78%) are available within both. Despite overlapping availability, most observations were primarily documented in only one of the two notes for each patient, with the exception of “Substance Use” and “Tobacco Usage” that were typically documented in both. For instances of dual documentation, there was a wide range of concordance in the value of the observation, ranging from about 20% to 100%, with the most commonly documented observations (e.g., marital status, living situation, substance use, and tobacco use) having concordance only about 75% of the time.

Conclusions
EHRs, including electronic clinical documentation, have been touted as a way to improve the quality of clinical care, including the ability to leverage data elements for quality and consistency. We have found that documented observations such as a patient’s social history is frequently inconsistently recorded by members of a patient’s health care team. Moreover, wide variation in the degree of concordance exists between notes authored during a single patient hospitalization. Future work should try to uncover reasons for differing documentation patterns, with the aim of streamlining and cross-verifying valued observations to promote quality, consistency, and accuracy of data. Current efforts must also be undertaken to align incongruous data in EHRs.

Tad Hirsch, PhD¹, Geoff Gray, BSc¹, James Gibson, MSc¹, Shrikanth Narayanan, PhD², Zac E. Imel, PhD³, David C. Atkins, PhD¹

¹ University of Washington, ² University of Southern California, ³ University of Utah

INTRODUCTION
Motivational Interviewing (MI) is a widely studied and effective method of psychotherapy for substance abuse treatment. However, the quality of psychotherapy in clinical practice remains largely unknown. In training contexts and clinical trials, the evaluation of psychotherapy providers relies on a human who listens to a recording of a session and makes decisions regarding whether certain therapist behaviors have occurred. Feedback can be provided informally (e.g., a supervisor listens to the session and offers their perspective and input) or in the form of standardized, numeric ratings from a theory-derived rating system – a process that has remained fundamentally unchanged for decades. Given the labor intensiveness of human ratings of psychotherapy, and the scale of demand in clinical practice, it is not surprising that most real-world psychotherapy is never evaluated. Thus, providers rarely receive feedback on the quality of their care. There is a strong need for an evaluation technology that makes feedback that is proximal to the clinical encounter feasible, which led to the development of the Counselor Observer Ratings Expert for MI (CORE-MI).

CORE-MI
CORE-MI combines speech and language processing with automated coding and interactive visualization to help therapists and trainers identify strengths and areas for improvement. CORE-MI utilizes foundational speech signal processing methods to translate an audio-recording of an MI session into a numeric representation of semantic and vocal acoustic data, which are in turn used as features in machine learning predictive models. Paralinguistic information such as prosody, pitch, speech rate, and intensity are also computed.

Figure 1. Core-MI user interface.

CORE-MI provides feedback on standard MI quality measures described in the Motivational Interviewing Treatment Integrity Scale. Two measures are considered “Global Ratings” and meant to capture the overall quality or “gestalt” of the session. The system also tracks three “Behavior Count” measures that characterize the quantity and quality of questions and reflections in the discourse. CORE-MI provides a report card-like, visual summary of counseling sessions (Figure 1). The report prominently features an “Overall MI Fidelity” score that aggregates all measures to give an impressionistic view of the participant’s general level of adherence to MI principles. Bar charts indicate global and behavior count measures, highlighting the counselor’s level of proficiency in MI techniques. The visualization also offers a “Session View” that enables users to examine detailed session transcripts and talk turn level annotations for therapist and client behaviors.

EVALUATION
Evaluation of the CORE-MI’s speech and language processing and automated coding achieved a correlation of 0.643 between expert annotated empathy codes and machine-derived estimations, and an accuracy of 81% in classifying high vs. low empathy [1]. A study is currently underway with therapists and therapists-in-training to evaluate utility, usability and desirability of the visualization system for clinical and training use. Results of this study are expected in late 2016.

CONCLUSION
The CORE-MI offers the first scalable technology for evaluating the quality of client-therapist interactions. The CORE-MI offers an additional contribution wherein feedback not only occurs rapidly, but is presented in a way that may facilitate therapist understanding and engagement with the raw material of the session. We are hopeful technological augmentations of psychotherapies similar to CORE-MI may lead to improvements in the quality of treatment available in the community and contribute to a reduction in the disease burden of substance abuse disorders.

REFERENCES
Ranking PubMed Hits by Projection on an Axis of Words Having Varied Relevance to Search Intention

Teruyoshi Hishiki, MD, PhD¹, Takuro Tamura²
¹Toho University, Funabashi, Japan; ²LINE Co., Ltd, Tokyo, Japan

Problem Addressed

In existing information retrieval models, bag-of-words queries represent the user’s search intention where the relevance of a document to the query is defined as the cosine of the angle between a word content vector made from the document and one from the query. We can consider the query vector as the axis onto which document vectors are projected. The scalar product of the query and document vectors gives the document a coordinate on the axis; the coordinate is proportional to the cosine value if the document length has been normalized. At this point, suppose that we introduce into the query another bag of words that have different meanings to those of the search intention and are assigned negative values. The coordinates of the document would then produce a spectrum ranging from the search intention to the different topic, and a cut-off for relevance could then be set easily and objectively.

Specific Purpose of the Project

The aim is to reduce the number of PubMed hits for a broad query to be scanned over by the user before gathering articles about the factors of decision errors in diagnosis. We chose aortic dissection (AD) as the search term because the symptoms associated with this condition overlap with those of other conditions, and some of the treatments for the other conditions are contraindicated for AD. We developed a browser-based interactive program that calculates the score for article titles as the sum of the word values in the title. The word values are defined by the user.

Interim Results of the Project

We obtained 5340 titles from a PubMed search for AD in titles. Four groups of words were indexed: symptom (24 words; value, 4), diagnostic modality (10 words; value, 2), causal or temporal (15 words; value, 1), and treatment or outcome (42 words; value, −3). The distribution of scores for all titles (Figure 1) shows large peaks for scores 14, 15 and 6, 7 (10 is the neutral score given to all titles) and smaller peaks for scores 18, 19 and 0, 1. We sampled the titles for them and found that most of the articles relevant to our search occurred within the intervals of scores higher than 13, with few less than eight. We estimated that ~10% of the titles needed further screening.

Conclusion

The addition of words with a negative value (i.e., with an “opposite” concept to the search intentions) into the query bag-of-words could facilitate making sense of information retrieval results. A limitation of this method is the effort required to find the opposite words. Prior knowledge as well as word frequency and co-occurrence information could be used to find these words.
Workflow-guided development of a clinical decision support tool for patients with advanced liver disease.

Samuel B. Ho, MD1,2, Julie E. Ducom, PhD3, Anne Miller, PhD3, Jennifer H. Garvin, MPH4,5, Jejo D. Koola, MD3,6, Russ Beebe, BS3, Jason M. Slagle, PhD3, Dax M. Westerman, MS5,6, Carrie Reale, MSN, RN3, Matthew B. Weinger, MD3,6, Erik J. Groessl, PhD1,2, and Michael E. Matheny, MD3,6

1 VA San Diego Healthcare System, San Diego, CA, 2 University of California, San Diego, CA 3 Vanderbilt University, Nashville, TN, 4 VA Salt Lake City Healthcare System, 5 University of Utah, Salt Lake City, UT 6 VA Tennessee Valley Healthcare System, Nashville, TN

Objectives: Patients with chronic liver disease and cirrhosis represent high-risk, high-cost patients, and optimizing the quality of care they receive is of high priority. We previously employed a Human Factors Engineering (HFE) approach to identify key design goals for clinical interventions aimed at improving the care of patients with cirrhosis1. Here, we describe the process of developing a clinical decision support (CDS) tool based on those design goals, with a focus on optimizing its integration into standard clinical workflow. Methods: A multi-site interdisciplinary collaborative Design Workshop consisting of 14 clinicians, graphic designers, informatics specialists, and health service researchers convened into two workgroups to develop a CDS tool prototype. Using participatory design, the two groups created user interface concepts in response to 8 structured scenarios of actual patients with chronic liver diseases in outpatient and inpatient settings. The groups’ designs were discussed and reconciled in a combined meeting. The resulting design concept included: 1) evidence-based order sets most relevant to patients at increased risk for liver disease; 2) an interactive web-based medium-fidelity prototype of the user interface (UI) design, consistent with the current electronic medical record (EMR) infrastructure; and 3) contextually relevant patient scenarios for multi-site usability evaluations of the UI prototype by front-line clinicians. These products were then iteratively refined based on structured interviews with primary care clinicians using web-based prototypes at three hospital sites. Results: The interactive UI prototype identifies patients at high risk for cirrhosis based on laboratory and radiology findings. Users are alerted to the patient’s condition via an automated alert on the existing EMR patient cover page. Selecting the alert leads to a web-based CDS split screen tool that provides the clinician with a consolidated view of risk factors and relevant test results on the left side. The right side pane gives users direct access to appropriate order sets, organized by clinical category. Actionable, recommended order sets are provided, with informational icons describing current evidence-based guidelines. Key elements refined by iterative development include different clinical categories, number and types of lab tests used, types of order buttons, and informational icons. Usability testing is underway. Conclusions: Following a HFE approach, we developed an interactive CDS tool to improve the risk stratification and care of patients with advanced liver disease, both in the acute care setting (e.g., decompensated cirrhosis) and for routine health maintenance. Following further refinement and full integration with the EMR, we plan to evaluate the impact of CDS use at two hospitals on provider acceptance, usability, and compliance with quality of care indicators for patients with cirrhosis.

Reference:
Development of an Electronic Trigger Tool for Identifying Inpatient Diagnostic Error

Katherine A. Homann, M.D., M.A.S. and Robert El-Kareh, M.D., M.S., M.P.H.

1University of California, San Diego, La Jolla, CA, USA

Introduction

Diagnostic errors—defined as incorrect or delayed diagnoses—are common and often preventable. Trigger tools have been proposed to improve the identification and study of diagnostic error; however, established criteria did not yet exist. We sought to create a trigger tool for diagnostic error in hospitalized patients through literature review and prioritization via established methods for expert review.

Methodology

Applying established methodology, the 29 diagnostic error trigger criteria identified by our comprehensive literature review underwent additional refinement. These criteria were reviewed and prioritized by two multi-disciplinary focus groups, a clinical advisory panel, and three Delphi rounds with experts in diagnostic error (Figure 1).

Results

The diagnostic trigger tool criteria were narrowed to 10 high yield triggers and 4 secondary tier triggers (Table 1).

Table 1. Final diagnostic error trigger tool criteria.

<table>
<thead>
<tr>
<th>Primary Diagnostic Error Trigger Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incident report for case that was classified as diagnostic error</td>
</tr>
<tr>
<td>Case referred to quality assurance or quality improvement committee</td>
</tr>
<tr>
<td>Complaint by patient or family related to speed or accuracy of diagnosis</td>
</tr>
<tr>
<td>Emergency department visit within 7 days before admission</td>
</tr>
<tr>
<td>Change of primary service within first 48 hours of hospitalization</td>
</tr>
<tr>
<td>Urgent or emergent surgery &gt; 12 hours after admission</td>
</tr>
<tr>
<td>Discrepancy between working diagnosis and (non-autopsy) pathology results</td>
</tr>
<tr>
<td>Discrepancy between working diagnosis and blood culture</td>
</tr>
<tr>
<td>Performance of autopsy</td>
</tr>
<tr>
<td>Newly abnormal tests which are not repeated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary Diagnostic Error Trigger Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased level of care</td>
</tr>
<tr>
<td>Subsequent readmission within 72 hours of discharge from hospital</td>
</tr>
<tr>
<td>Discrepancy between admission diagnosis and discharge diagnosis</td>
</tr>
<tr>
<td>Discrepancy between preliminary and final radiology diagnoses</td>
</tr>
</tbody>
</table>

Figure 1. Methods for trigger tool creation

Conclusions

Diagnostic error is a significant source of preventable healthcare harm. Our study established a preliminary trigger tool for diagnostic error in hospitalized patients. These criteria should be prospectively validated in different settings to verify their utility in identifying important diagnostic errors.

References

Perceived Effectiveness of Clinical Decision Support to Improve VTE Prophylaxis: A Comparison of Physicians, Pharmacists and Nurses

Peter Hoonakker1, PhD, Shimeng Du1,2, BSc, Ann S. Hundt1, PhD, Jason Stamm3, MD, Brian Patterson, MD4, Pascale Carayon1,2, PhD
1 Center for Quality and Productivity Improvement, University of Wisconsin-Madison
2 Department of Industrial and Systems Engineering, University of Wisconsin-Madison
3 Department of Medicine, Geisinger Health System, Danville, PA
4 Department of Emergency Medicine, University of Wisconsin-Madison

Venous thromboembolism (VTE) is a frequent, but preventable complication for hospitalized patients that can lead to significant morbidity and mortality [1-3]. It has been estimated that more than 900,000 VTE cases occur in the US every year [4]. In our study, we use a sociotechnical systems approach to determine design requirements for clinical decision support (CDS) to prevent VTE and address some of the cognitive and team work issues that have limited the adoption, acceptance and effectiveness of CDS for VTE. To better understand the socio-technical context of the VTE prophylaxis process we asked the principal clinicians in the process (physicians, pharmacists and nurses) about their perceptions of VTE prophylaxis and the perceived effectiveness of different CDS interventions to improve the process, including computer, human and organizational decision support.

We conducted a survey on perceptions of clinicians with regard to importance, effectiveness, safety and utilization of VTE prophylaxis and on possible effectiveness of interventions to address VTE prophylaxis among more than 700 clinicians (physicians (n=189), pharmacists (n=56) and nurses (n=456) in two hospitals (response 92%).

Results show that, overall, the three clinician groups think that VTE prophylaxis is important, safe, effective and appropriately utilized. With regard to possible interventions to improve VTE prophylaxis, results show that all respondents have a strong preference for computer CDS. Out of the 16 possible interventions, the 4 top-rated interventions for all clinician groups are technological interventions: 1. computerized alerts to inform that prophylaxis has been interrupted and needs to be resumed, 2. order sets, 3. a hard stop in the order set, and 4. an alert that contra-indications have changed. Note that some of these interventions involve smart alerts. The first human intervention (pharmacist reminders) ranks 5th in preference by physicians and 8th by pharmacists and nurses. Organizational interventions such as educational meetings, periodic audits, feedback, posters on the units and laminated pocket cards as reminders were preferred least. Unlike pharmacists and nurses, physicians have a strong preference for an alert that informs them that prophylaxis has been interrupted and needs to be resumed. Otherwise differences in preferences between clinicians are relatively small.

VTE prophylaxis is not limited to a single task (e.g. placing an order for prophylaxis) but is a complex process that continues to be addressed throughout the patient stay (e.g. resuming VTE prophylaxis after a procedure), in which multiple actors are involved. The literature [4] shows that a single intervention (e.g. education and training) will not improve VTE prophylaxis significantly. What is needed is CDS that supports all clinicians throughout the hospital stay that is not limited to health IT, and takes teamwork into account. A socio-technical approach can help to identify the factors that can improve VTE prophylaxis in this context.

References
Implementing Semantics-Based Cross-domain Collaboration Recommendation in Biomedicine with a Graph Database

Dimitar Hristovski, PhD¹, Andrej Kastrin², Thomas C. Rindflesch, PhD³
¹Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia; ²Faculty of Information Studies, Novo mesto, Slovenia; ³National Library of Medicine, Bethesda, MD, USA

Introduction
Research collaboration is an important condition for conducting high quality science. Current systems provide a list of potential collaborators but do not further motivate the basis for the recommendation. We address this deficiency and propose a novel approach for semantics-based recommendation of research collaboration that exploits an extended literature-based discovery (LBD) methodology. The system provides a list of potential collaborators and topics for collaboration, and also explains why the collaboration is compelling⁴. Here we describe the implementation of this methodology with the Neo4j graph database and its query language Cypher².

Methods and Results
We first constructed a large network and loaded it into the Neo4j graph database. The network consists of two major types of nodes: authors and biomedical concepts. We extracted authors from the full MEDLINE bibliographic database. We used SemRep³ to identify biomedical concepts from the set of arguments (subjects or objects) of semantic relations extracted from all MEDLINE titles and abstracts. Our network contains several types of arcs and edges. co_author edges link any two authors who have been co-authors on at least one paper. writes_about arcs link authors to biomedical concepts. These arcs are derived from the semantic relations extracted from articles written by the authors. The writes_about arcs are used to represent the expertise of the authors. Finally, we have 30 types of semantic relations extracted with SemRep that link the nodes representing biomedical concepts. These relations represent current biomedical knowledge. The network consists of: 9,516,106 author nodes, 269,047 biomedical concept nodes, 181,664,746 co_author edges, 189,294,999 writes_about arcs, and 69,333,420 arcs that represent semantic relations between biomedical concepts.

We implemented the algorithm for recommending research collaboration with the Cypher query language and it operates as follows. For a given input author, we first compile the author’s topic (concept) profile, which represents both the authors interests and expertise; this is done by following the writes_about arcs as described above. For each input concept we perform an open LBD discovery. LBD proposes target concepts as novel collaboration (research) topics that are not yet published in the literature. For all target concepts found by LBD, we find authors who have these concepts in their profiles and eliminate those authors who are already coauthors with the starting author. The output is a list of the remaining authors as potential collaborators along with topic(s) for collaboration. Shown below is a generic implementation with a Cypher query, which can be made more specific as needed:

MATCH (author1:author) -[:WRITES_ABOUT]-> (X:Concept) -[Rel_XY]-> (Y:Concept) -[Rel_YZ]-> (Z:Concept) <-[:WRITES_ABOUT]- (author2:author)
WHERE NOT (X)-[RelXZ]->(Z) AND NOT (author1)-[CO_AUTHOR]-(author2)
RETURN author1, X, Rel_XY, Y, Rel_YZ, Z, author2;

Conclusion
Using a graph database such as Neo4j for storing the large network data structure needed for semantics-based cross-domain collaboration recommendation is more natural and efficient than using a relational database. Implementing collaboration recommendation algorithms is conceptually easier and more simple when using a graph query language such as Cypher when compared to standard SQL.

References
2. Neo4j graph database: http://neo4j.com/
Machine Learning Methods for the Early Identification of Diabetes

Kang Lin Hsieh, MS¹, Susan.H.Fenton,Ph.D¹
¹University of Texas Health Science Center School of Biomedical Informatics

Introduction: Identifying pre-diabetes and undiagnosed diabetes are needed for the prevention of diabetes. Two major tests, the hemoglobinA1c (HbA1c) and fasting plasma glucose (FPG), are used to diagnose pre-diabetes and diabetes. However, persons with diabetes or this predisposition do not always have consistent results between these 2 tests. Further study of the patients with discordant and concordant test results may assist in identifying new relevant risk factors leading to pre-diabetes and undiagnosed diabetes. [1]

Material and Methods: The National Health and Nutrition Examination Survey is a biannual survey that enrolled 78,518 people from 1999-2011. Variables included in the predictive model for the development of diabetes included daily eating habits, interview of diabetic risk factors, demographic data, and lab tests relevant to diabetes. The demographic data included age group, not age for anonymity purpose. 7,915 records were selected for this research. In this population, 5,377 people had concordant results between their HbA1c and FPG test and 2,538 people had discordant results. A prediction model for the development of diabetes was created using the concordant data and tested using the discordant group data to determine the impact of the lab tests on predictive accuracy.

Results: The prediction results between the 2 groups were not significantly different for any factors except age. In Figure 1., the total population predictive accuracy was 61.4% for the HbA1c and 31.4% for the FPG. When examined by age group, the over 65 accuracy was 39% for the HbA1c and 57% for the FPG, age 40-65 was 57% for the HbA1C and 34% for the FPG, age 18-40 was 77% HbA1C and 20% FPG, while age less than 18 was 79% HbA1C and 20% FPG respectively.

Figure 1. RandomForest Analysis Results

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>HbA1c, 79%</th>
<th>FPG, 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>61.90%</td>
<td>57%</td>
<td>77%</td>
</tr>
<tr>
<td>31.40%</td>
<td>39%</td>
<td>20%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>over 65</th>
<th>40-65</th>
<th>18-40</th>
<th>below18</th>
</tr>
</thead>
<tbody>
<tr>
<td>All population</td>
<td>HbA1c</td>
<td>FPG</td>
<td>HbA1c</td>
<td>FPG</td>
</tr>
</tbody>
</table>

Figure 2. Material of research

Discussion: The model based on the concordant group did not achieve high predictive accuracy in the overall population. In further analysis, the accuracy of HbA1c is usually higher than FPG except in the group aged 65 and over. These results suggest the value of the HbA1c in predicting diabetes is more dominant than the value of the FPG except for those aged 65 and over. This is a very interesting phenomenon. Confounding factors or unknown relationships may also impact this analysis.

Conclusion: The conclusion is that the two populations are very different especially above age 40 which is when the risks of being diagnosed with diabetes increases dramatically. Continued study is needed to identify the factors resulting in the discordance between the HbA1c and FPG. This can provide a new perspective on identifying pre-diabetes and undiagnosed diabetes.

References

A mockup interface for searching similar patients using temporal patterns

Henry Huang¹, Xiaoqian Jiang, PhD¹

¹Health System Department of Biomedical Informatics, UC San Diego, La Jolla, CA 92093

Abstract
We build a mockup interface for a patient to query statistics of similar patients based on his or her symptoms and medical event history. This interface is a part of an ongoing deep learning project of developing predictive models for early prognosis. Treating patient reported data as a timeline, we can generate the statistics of similar patients. Available online: http://dquery.ucsd-dbmi.org:8800/formssets/

Introduction
Patients have benefited from online knowledge bases like WebMD [1] or online patient networks such as PatientsLikeMe [2]. But it is not easy to find the right information for an individual given the complexity of medical conditions. It would be more useful to have a web service that aligns temporal patterns of symptoms and medical events of patients in order to provide users with useful information (e.g., statistics of diagnosis and medication of similar patients) that is most beneficial to their own personal health.

Methods
Our predictive model uses a vector space model to represent each medical event in a continuous vector space. Construction of the vector space with training data take into account each medical event’s neighbors in the timeline with a 2-layer neural network. We were able to achieve AUC of average of 78% in predicting over 30 diseases. This predictive model paper is currently under review. Using this model in the backend, users can enter their symptoms, diagnosis, treatment, conditions, and/or symptoms into our interface in a chronological order. Given such sequence of events, the backend predictive model will project the user onto the vector space in order to find most similar patients in the same vector space and return the next most likely diagnoses and treatments, as well as corresponding statistics (as illustrated in the figure below).

References
Evaluating Patients’ Perceptions of Infobuttons

Nathan C. Hulse, PhD1,2, Jie Long, PhD1
1Intermountain Healthcare, Salt Lake City, UT; 2Department of Biomedical Informatics, University of Utah, Salt Lake City, UT;

Abstract

Infobuttons have been used in electronic health records for some time, and their value in providing context-aware information for clinicians has been demonstrated by previous research. However, their utility in providing answers for patients directly via the personal health record has not been well characterized. In this poster we will present results of an experiment in which patients are presented with infobuttons related to problem list, medication list, and lab value entries, and asked to evaluate the relevance of the results returned. We will present descriptive data and agreement statistics that describe the perceived relevance of the returned results for a number of information providers. We will also present survey data related to patients’ perceptions of infobutton usefulness and ease of use.

Introduction

One of the important challenges of the information age is that of quickly providing access to relevant, well-sourced material that can answer questions well at the time and place that they arise. In medical informatics, infobuttons have been shown to be an effective tool in meeting this challenge for clinicians, yet, the same types of resources targeted directly for patients have not been well-characterized.

Background

In 2014, we implemented a local instance of the OpenInfobutton project and made it available for patients to use from our patient portal, myHealth. We enabled seven different electronic resource providers from five different domains in the patient portal, including problems, medications, lab results, allergies, and microbiology. In order to study patients’ impressions of the relevance and usefulness of the results returned by the infobutton, we conducted a pilot study to quantify patients’ perceptions of the tool.

Methods

We sampled a set of concepts from a commonly used panel of data derived from recent EHR usage metrics; identifying 15 each (10 common, 5 less so) from labs, medications, and problems and created infobuttons for patients to test. We asked 5 patients to review and score each results for relevance, and usefulness, as well as complete a survey of their perceptions of the tool’s ease of use and usability. Relevance and usefulness scores ranged from 0-3, per a qualitative scale, while the remaining survey questions were based on the TAM model.

Results

Statistical analysis showed clear differences across information providers as to overall relevance and utility of results returned. Patient-facing infobuttons performed better in medication list than in lab value entries and problem list. Providers who offer infobutton content through standardized HL7 APIs consistently scored higher for relevance and utility. Overall, survey data showed patients’ perceptions of infobuttons as highly useful and easy to use.

Conclusion

We have evaluated seven different resources linked from patient-facing infobuttons to test patients’ perspectives about their relevance, usability, and usefulness. In doing so, we identified several resources which perform better than their counterparts in providing relevant results; the data showed significant differences across content providers within domains. The value of HL7-based access to content for providing relevant content was evident in the results. TAM-based survey data suggests that patients find infobuttons both usable and useful. We feel that our focus on understanding patients’ information needs will empower them as they use consumer-facing informatics tools.

References

Triggers for VTE Prophylaxis: Sociotechnical System Considerations

Ann S. Hundt, PhD1, Peter Hoonakker, PhD1, Jason Stamm, MD2, Vaibhav Agrawal, MD2, Brian Patterson, MD3, Pascale Carayon, PhD1,4

1Center for Quality & Productivity Improvement, University of Wisconsin-Madison
2Dept of Medicine, Geisinger Health System, Danville, PA
3Dept of Emergency Medicine, University of Wisconsin School of Medicine & Public Health
4Dept of Industrial & Systems Engineering, University of Wisconsin-Madison

It is estimated that more than 900,000 venous thromboembolism (VTE) cases occur in the US every year1 and about sixty-percent of the people with VTE developed it during, or within 30 days of, their hospitalization.1,2 In addition to health consequences, VTE represents a major economic burden attributed to the costs of the initial hospitalization and readmissions due to VTE (5-14%).3 Prophylaxis is known to decrease the incidence of VTE. Therefore VTE prophylaxis has received significant attention from physician professional societies (e.g., American College of Chest Physicians) and the Joint Commission. In the study described here, we gained a greater understanding of triggers, in light of the sociotechnical system, that affect VTE prophylaxis ordering decisions throughout patient hospitalizations.

In this study we observed 8 morning rounds on 3 services for a total of 13 1/2 hours, focusing on clinical practice and discussion related to VTE prophylaxis orders. We also conducted 18 interviews of attending physicians and residents lasting a total of 12 hours and 13 minutes. In the interviews we addressed VTE prophylaxis ordering practices and communication at six stages: on patient admission, during daily rounds, when prophylaxis is interrupted, when prophylaxis warrants re-initiation, cases when prophylaxis is not ordered on admission due to a contraindication that later stabilizes or is resolved, and on patient transfer from a different clinical service. We analyzed the observations in which VTE prophylaxis communication occurred and conducted qualitative content analysis of the interviews and identified 1) the initiator and others participating in VTE prophylaxis-related communication, 2) the technology and/or tools used to facilitate the communication, 3) whether the communication occurred within or outside of team rounds and 4) clinical practice (idiosyncratic and not) that influenced ordering. We then identified the triggers that prompt such communication and changes to existing VTE prophylaxis orders.

Aside from clinical information in the electronic health record (EHR) that influences prophylaxis ordering (e.g., lab values denoting heparin-induced thrombocytopenia, renal insufficiency, patients scheduled for procedures such as pain control with an epidural anesthetic), other aspects of the sociotechnical system provided triggers that influenced the decision to continue, stop or (re)initiate prophylaxis. In many instances, verbal communication outside rounds prompted timely changes in prophylaxis orders. This was especially true of nurses who, though they generally participate in rounds, are more influential outside of rounds when they identify the need to stop or re-initiate prophylaxis. Preferences by proceduralists and their willingness to perform procedures on patients recently (or currently) receiving anticoagulants also influenced orders to hold, stop or continue VTE prophylaxis. These findings point to the need to understand the sociotechnical system in which clinical teams function and communicate, and to determine the role of the EHR in providing meaningful, timely clinical decision support for VTE prophylaxis.

Acknowledgments
This project was supported by grant number R01HS022086 from the Agency for Healthcare Research and Quality and the CTSA program, through the NIH NCATS, grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ or NIH.

References
Predicting ordered diagnostic tests from patient triage data

Haley Hunter-Zinck, PhD1, Stephan Gaehde, MD, MPH1
1 VA Boston Healthcare System, Boston, MA

Introduction

Emergency departments (EDs) are continuously working to increase patient satisfaction and reduce length of stay. Laboratory testing or imaging procedures are often ordered only after evaluation of the patient by a provider. There is an opportunity to improve ED efficiency by shortening the interval of time between triage and diagnostic testing. Accurate prediction of appropriate complaint specific diagnostic testing has the potential to allow ordering to be initiated immediately after patient triage rather than later in the visit when the provider conducts an initial patient assessment. With predicted, triage-time ordering, diagnostic test results could potentially be available to providers during this initial patient assessment, allowing them to initiate patient treatment immediately. To pursue this aim, we investigated whether we could predict a patient’s ordered tests from data collected at triage.

Methods

Using the National Hospital Ambulatory Medical Care Survey, a publicly available dataset from the Centers for Disease Control and Prevention, we extracted information on patient visits that would be available upon triage or from previous medical history as well as procedures ordered during the visit. Using a multi-label machine learning framework, we compared performance of binary relevance and label powerset methods, the Random k-labelsets method, and a modified multi-label neural network classifier. We assessed prediction performance in aggregate and for each order individually as well as calculating the relative importance of each data feature.

Results

Prediction performance varied greatly depending on the test but mostly due to its frequency of administration in the dataset. Performance, as measured by the F1-score, as well as the frequency with which each test was ordered, is shown in Figure 1. For example, we predicted the order of a complete blood count, administered in 44% of sampled visits, with a F1-score of 0.74 while we predicted blood alcohol assessments, administered in less than 4% of sampled visits, with a F1-score of 0.43. Several variables were important for prediction across all procedures, including arrival by ambulance, acuity score, age, and injury.

Conclusion

Overall, we have adequate information in triage data alone to predict relatively common test orders. Variables important for prediction are standard data collected at triage or readily available in the electronic health record.

Figure 1. (left) Frequency of diagnostic blood test orders over all sampled visits. (right) Classification performance using a multi-label neural network approach as measured by the F1-score for diagnostic blood test orders.
Evaluating Differences Between MIMIC II and III Critical Care Databases

Matias I. Hurtado, Undergraduate1, Jette Henderson, Phd Student2, Joydeep Ghosh, Phd2
1Pontificia Universidad Católica de Chile, Santiago, Chile; 2The University of Texas at Austin, Austin, Texas, USA

Introduction The Medical Information Mart for Intensive Care (MIMIC) is an Electronic Health Record (EHR) database containing information about patient visits to the Beth Israel Deaconess Medical Center Intensive Care Unit (ICU). MIMIC has had three versions to date. MIMIC II contains ICU patient information from 2001 to 2008, and MIMIC III, which was released in 2015, extends MIMIC II with data from 2008 to 2012.

In this poster, we begin to explore the differences between MIMIC II and III using the following approaches: 1) compare MIMIC II to III using descriptive statistics; 2) perform topic modeling (LDA) on MIMIC II and III to examine whether latent topics change significantly between the two datasets; and 3) create a mortality prediction modeling using structured and unstructured (mostly nurses notes) data. The third approach closely follows the methodology of Ghassemi et al.,2,3 which was successfully applied to MIMIC II. Our results provide several indications regarding the value of the newer dataset and changes in the population that the data represents and will be valuable specially to the large research community that has previously worked on MIMIC II.

Methods Observations and descriptive statistics were extracted from PostgreSQL databases housing MIMIC IIv1.2 and IIIv1.3 respectively. We followed Ghassemi et al. and first filtered patients using the following criteria: patients must be older than 18 years old and patients must have more than 100 non-stopwords in their notes. This filtering led to 19,792 patients with 25,026 ICU stays in MIMIC III and 14,656 patients with 17,079 ICU stays in MIMIC II. We explored the differences by applying Latent Dirichlet Allocation (LDA) to clinical notes of eligible patients in MIMIC II and III. A vocabulary for the LDA models consisted of the 500 most informative words of each patient based on a tf-idf model.

For the mortality prediction models, we used several distinct feature sets, but for the sake of brevity, we only discuss the Admission Baseline Model here, which contains basic features such as gender, age, and SAPS-II admission score. These features serve as the input for SVM models that use the radial basis function as a kernel function to cater to non-linear effects.

Results We discuss a subset of results here, and will show a more complete set in the poster. Using the Admission Baseline Model on MIMIC III we obtained an AUC of 0.93, in contrast to the earlier2 MIMIC II study, which obtained an AUC of 0.77. Examining the features individually reveals that this high score is due to the SAPS-II score feature.

The plot shows the median proportion of topics by survival outcome for MIMIC III. As Ghassemi et al. found, some topics seem to be correlated with the likelihood of long-term survival.

Conclusions One important step to harnessing the power of MIMIC III is exploring the differences between MIMIC II and III. In this poster, we begin that process by replicating a MIMIC II study performed by Ghassemi et al. using MIMIC III as well as explore the latent features of MIMIC III.

References
Exploring the Use of ClinicalTrials.gov Trial Results Data for Pharmacovigilance

Vojtech Huser, Olivier Bodenreider
Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction
Pharmacovigilance aims to monitor drug safety using sources such as spontaneous reporting systems, biomedical literature or electronic health record data. Clinical trials represent a source of drug-event pairs data complementary to these sources for signal detection in pharmacovigilance platforms. The advantage of ClinicalTrials.gov (CTG) over other pharmacovigilance sources is the large number of negative drug-event pairs (explicit evidence that a given drug is not causing a particular adverse event; count of 0 is reported in the deposited results). With 208,959 trials registered and 20,025 trial results, CTG is the largest repository of trial summary results (with more than 3700 new trial results deposited per year). CTG trial registration data provide information about trial type, sponsor, arms and interventions. CTG results data further provide trial participant counts, baseline characteristics, outcome measures and, most importantly for our study, significant adverse events. Adverse events (AE) are recorded separately by trial arm (or trial group). Although some data submitted to CTG are structured, such as number of arms or intervention type (e.g., drug vs. procedure), many elements are collected as free text (e.g., the drug(s) used in the trial). This preliminary investigation explores the selection of clinical trials of interest for pharmacovigilance and the feasibility of extracting drug concepts from CTG trial registration data.

Methods
To investigate the proportion of drug trials that can be directly used for pharmacovigilance (without additional manual curation), we analyzed results of interventional trials with drug interventions. We used CTG’s tabular data format and the structured element intervention_type. To identify drug names in the CTG intervention_name field, we mapped them to RxNorm using increasingly aggressive techniques, namely using the findRxcuiById (exact/normalized match) and getApproximateMatch functions of the RxNorm API.

Preliminary Results
Feasibility counts: As of February 22, 2016, the tabular CTG data included a total of 14,007 results of interventional trials that had at least one drug intervention (dataset S1; supplemental data are available at github.com/vojtechhuser/CTG). We found that 5,192 trials (37% of our sample) have exactly one trial arm and drugs from such trials can be unambiguously associated with AEs reported in the deposited trial summary results. Additionally, in 2,049 two-arm-trials (14.6% of our sample) we converted their free-text specified placebo arms into formally modelled placebo arms which may allow us to use additional pharmacovigilance methods. Overall, a total of 7241 trials would be amenable to processing for pharmacovigilance purposes.

Drug term detection: We processed 63,817 unique interventions strings (extracted from all CTG interventions of type ‘drug’). For 10.4% of those strings, the RxNorm API exact or normalized match function identified RxNorm concepts of type ingredient, clinical drug or branded drug (dataset S2). When no exact or normalized match was found, we proceeded with approximate match and were able to map additional 0.6% of input strings without the need for human review (single RxCUI detected with 100% detection certainty score; dataset S3). For the remaining input strings, approximate match provided multiple inputs (6.1 RxNorm terms on average) with a wide range of scores.

Conclusions
These results indicate that intervention names are usually not simple drug names and would require parsing for extraction of drug names (e.g., with MedEx, a medication text processing tool). Overall, our work indicates that a significant portion of the CTG result database can generate some drug-adverse event pairs that can be used for pharmacovigilance purposes. In our experience, the main obstacle to leveraging CTG for pharmacovigilance is the difficulty in unambiguously associating information from registration (intervention drug + trial arms) with information from result summaries (AEs), because of the absence of an explicit link between trial arms in these two resources.

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.
Secondary Data Analysis on Ventilator-associated Pneumonia and Pressure Ulcer with Regards to Head of Bed Elevation

Sookyung Hyun, RN, PhD1, Cheryl Newton, RN, 2
Susan Moffatt-Bruce, MD, PhD3

1Pusan National University College of Nursing, Busan, Korea, Wexner Medical Center at The Ohio State University 2Department of Critical Care Nursing, and 3Chief Quality and Patient Safety Officer, Department of Surgery, Columbus, OH, USA

Introduction

Typical process measures to prevent ventilator-associated pneumonia (VAP) are elevation of the head of the bed (HOB) as lower HOB is associated with incidence of VAP. Conversely, pressure ulcer (PU) prevention guidelines recommend that HOB positioning should be lower to reduce risk for PU development which contradicts VAP prevention guidelines for the HOB between 30 and 45 degrees for intensive care unit (ICU) patients. This presents a care dilemma and tension. The purpose of this study was to perform a secondary data analysis using cumulative electronic health record data in order to determine the association of HOB elevation with VAP and pressure ulcer development in ICU patients.

Methods and Results

Data were retrieved from an information warehouse from an academic medical center. The dataset was from adult ICU patients who were admitted to adult ICUs between January 1st, 2007 and December 31st, 2010 and they were on ventilator during their ICU stay. Pressure ulcer and VAP data were recorded in physicians discharge documentation and HOB data were recorded in nursing documentation, respectively, in the electronic health record systems. Descriptive statistics and chi-square test were used to analyze the data. Table 1 summarizes characteristics of patients and the association was illustrated in Table 2.

Table 1. Characteristics of Patients (N=7739)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>4432</td>
<td>57.3</td>
</tr>
<tr>
<td>Female</td>
<td>3307</td>
<td>42.7</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>6343</td>
<td>82</td>
</tr>
<tr>
<td>Black</td>
<td>1087</td>
<td>14</td>
</tr>
<tr>
<td>Hispanic</td>
<td>57</td>
<td>0.7</td>
</tr>
<tr>
<td>Asian/native Hawaii</td>
<td>48</td>
<td>0.7</td>
</tr>
<tr>
<td>American Indian</td>
<td>14</td>
<td>0.2</td>
</tr>
<tr>
<td>Other</td>
<td>81</td>
<td>1.0</td>
</tr>
<tr>
<td>Mean Age</td>
<td>57.8</td>
<td>15.9</td>
</tr>
<tr>
<td>Length of ICU stay</td>
<td>10.1</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Note. *=N is less than 7739 due to missing data

Table 2. Frequencies of VAP and pressure ulcer by HOB

<table>
<thead>
<tr>
<th>HOB</th>
<th>VAP Yes</th>
<th>VAP No</th>
<th>Pressure Ulcer Yes</th>
<th>Pressure Ulcer No</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 30°</td>
<td>325</td>
<td>7074</td>
<td>574</td>
<td>6825</td>
</tr>
<tr>
<td>&lt; 30°</td>
<td>8</td>
<td>332</td>
<td>10</td>
<td>330</td>
</tr>
</tbody>
</table>

Note. VAP=ventilator-associated pneumonia; HOB=head of bed

The association between HOB elevation and pressure ulcer occurrence was significant, $\chi^2(1, N=7739)=10.8$, p=.001 while the association between HOB elevation and VAP was not significant, $\chi^2(1, N=7739)=3.28$, p=.07.

Discussion

In our ICU care setting, patients regularly had their HOB greater than 30 degrees with pressure ulcer prevention measures. Nurses documented HOB every 2 hours or more frequently. Our data showed that HOB elevation was not likely to be associated with VAP whereas it was likely to be related to pressure ulcer occurrence. The findings of this study have limited generalizability and additional research is needed to better understand risks and benefits of process measures, including HOB elevation, to prevent ICU associated safety events.

Acknowledgements

The project was supported by UL1RR025755 from the National Center For Research Resources and a 2-Year Research Grant of Pusan National University.
Assessing the Feasibility of Developing an Inpatient Telehealth Referral System for Veterans with Poorly Controlled Diabetes

Umar Iqbal, MD¹, Khaleel S. Hussaini, PhD², J Edward Maddela, MD³, Hamed Abbazadehgan, MD, MBA⁴
¹,²,³University of Arizona College of Medicine, Phoenix, Arizona; ⁴Phoenix VA Healthcare System, Phoenix, Arizona

Introduction
The Phoenix VA Healthcare System presently has various home telehealth programs that are offered to veterans if they meet eligibility criteria. Presently, veterans are enrolled into these programs from physician referrals in the ambulatory clinics. There are no referrals being placed from the inpatient setting. “Telemedicine-based care diabetes management improves outcomes versus clinic care but is seldom implemented by healthcare systems”¹. In a recent study by Crowley et al., baseline HbA1c was at 10.5%, and by 6 months, estimated HbA1c had improved by 1.3% for intervention participants with telehealth visits and 0.3% for usual care¹.

Objective/Aim
To assess the feasibility of implementing an inpatient telehealth referral system prior to scaling up, we retrospectively examined baseline HbA1c values for veterans who are currently enrolled in the diabetes telehealth program from the ambulatory setting at the Phoenix VA Healthcare System.

Methods
This single-group pretest posttest study utilized baseline and current HbA1c data from enrolled veterans in the diabetes telehealth program from an ambulatory setting during August 2015 to December 2015. Inclusion criteria for the diabetes telehealth program requires a documented HbA1c ≥ 8.0, veteran’s to be on insulin, and are willing and able to perform finger sticks as instructed. A total of 169 veteran’s baseline HbA1c and most current HbA1c’s were extracted from the CPRS system who were enrolled in the telehealth program. Univariate, bivariate, and multivariate statistical procedures were conducted using SAS v9.2 (SAS Institute, Cary, NC). Univariate analyses included examining the distribution of the HbA1c and appropriate transformations were performed to reduce skewness and normalize the data. We assessed differences in baseline HbA1c by key demographic variables and used chi-square test for nominal variables and t-tests for continuous variables. We performed a paired t-test to assess differences in baseline and current HbA1c and estimated the effect size as small, moderate, and/or large based on Cohen’s criteria.

Results
The average age of the currently enrolled veterans was 64.63 years (SD = 8.80) and approximately 95 percent of the veterans were males (n = 160), and slightly over 50 percent of them were married and the majority of them resided in an urban county (~95%). The current HbA1C values (M =8.42; SD = 1.50) were lower as compared to baseline values (M = 9.15; SD = 1.20) t(168) = -7.60, p < .0001, d = -0.53 with a moderate effect.

Discussion and Conclusion
Preliminary evidence suggests that the veteran’s currently enrolled in the diabetes home telehealth program had a lower Hb1Ac value after being enrolled into the program. While the impetus to scale up the new inpatient telehealth program is promising, the study is limited due to lack of randomization, availability of a proper control group and perhaps subject to selection bias. In addition, the study did not account for factors influencing utilization of the current telehealth program from the ambulatory setting. This new pilot inpatient telehealth referral program, that is currently underway, bolsters some of these limitations by assessing factors that influence utilization of the telehealth program and developing clinical decision support to alert inpatient program coordinators to enhance utilization.

References
Information and Communication Needs of Hispanic Dementia Caregivers
Informing Design of a Family Health Information Management System

Sarah J. Iribarren, RN, PhD1, Samantha Stonbraker, MPH, RN1, Niurka Suero-Tejeda, MS, MA, CHES,1 Robert Lucero, RN, MPH, PhD, FAAN2,3, Suzanne Bakken, RN, PhD1
1Columbia University, New York, US; 2University of Florida, Gainesville, FL, US; 3VA HSR&D Center of Innovation on Disability and Rehabilitation Research

Abstract
This study used a multimethod qualitative design to elucidate information and communication needs of Hispanic caregivers for those with dementia and required functionalities for an online information management system.

Introduction
Dementia is an increasing problem and is more prevalent in Hispanics.1 Caregivers for patients with dementia are at high risk for poor mental and physical health outcomes.2 Little is known about the information and communication needs of Hispanic caregivers for individuals with dementia or the manner in which online tools may meet their needs.

Methods
We conducted 11 participatory design sessions with 10 English- and 14 Spanish-speaking Hispanic caregivers to understand their information and communication needs and potential tools to meet these needs. Audio/video recordings of sessions were transcribed verbatim and analyzed using a theory-based framework3 with three-level coding in Nvivo 10. Rigor and replicability were supported through member checks during sessions, peer debriefing following sessions, coding by multiple coders, documenting an audit trail, and meetings to resolve differences.

Results
Participants were 60 years old (±7.7), caregivers for an average 6.5 years (±4.2), and spent an average of 68.7 hours caregiving per week (±56.6). The majority were female (80%), spoke Spanish as the primary language at home (67%), and were caregiving for a parent with dementia (67%). Communication and Information needs or tasks were reported by the majority of participants (21 and 22, 87.5% and 91.7% respectively) while fewer reported online tools (19, 79.2%). Not knowing how to care for their ‘loved one’ effectively and needing reassurance as to what to do were common themes. Results included:

- Communication needs: communicate with healthcare team/provider for questions as they arise, ability to notify others when the person with dementia is lost (alerts), and communicate with other family caregivers.
- Information needs: information on dementia type and progression, “how to” information (e.g., lifting techniques, better take care of themselves), about medications (e.g., effects, side effects and drug-drug interactions), community services, and health status of individual (e.g., lab results, medical records).
- Online tool needs for: care coordination (e.g., appointment tracking, find/communicate with healthcare providers), important information (e.g., phone numbers, resources, medications, immunization records), support groups, missing person alert, and schedule or to-do list that can be marked off as completed.

Conclusion
Family caregivers have a range of needs related to self-managing their role that inform the FHIMS design.

Acknowledgment: Funded by New York City Hispanic Dementia Caregiver Research Program. R01NR014430.

References
Understanding Delays In Abnormal Test Result Follow-Up Using Electronic Health Records In Outpatient Primary Care Settings

Roosan Islam\textsuperscript{1,2}, PharmD, PhD Viraj Bhise, MBBS, MPH\textsuperscript{2,3}, Janet Schwartz-Micheaux\textsuperscript{1,2}, Elise Russo, MPH\textsuperscript{1,2}, Daniel R. Murphy, MD, MBA\textsuperscript{1,2}, Dean F. Sittig\textsuperscript{1,3}, PhD, Hardeep Singh\textsuperscript{1,2}, MD, MPH

\textsuperscript{1}Department of Medicine, Baylor College of Medicine, Houston, TX
\textsuperscript{2}Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX
\textsuperscript{3}The University of Texas Health Science Center, Houston, TX

Background:
Delays in follow-up of abnormal test results are a significant safety concern in outpatient settings. Although, electronic health records (EHRs) can help ensure reliable delivery of important clinical information, they do not guarantee appropriate follow-up actions are taken. In order to improve test results follow-up, it is important to understand how clinicians’ follow-up abnormal test results within the complex “socio-technical” context of EHR-enabled healthcare systems. In this study, we sought to explore reasons for delays in abnormal test results follow-up.

Methods:
We recruited primary care physicians from multiple family medicine outpatient clinics site in Houston, TX. We queried the site’s clinical data repository from January 1\textsuperscript{st}, 2015 to September 30\textsuperscript{th}, 2015 to identify potential delays in abnormal results follow-up for ten common imaging, laboratory and pathology tests (e.g. Hemoglobin, Thyroid-Stimulating Hormones, Chest X-Ray, PAP Smears) within patient charts. A physician reviewed these selected charts to evaluate if appropriate follow-up took place within 14 days of the patient visit. We defined a delay in follow-up if there was no documentation of patient notification, repeat or follow-up testing, referral or change in medications within 2 weeks of the test result. In this ongoing study, we have thus far interviewed 5 physicians in which a delay was identified. We interviewed participants using the critical decision method (CDM), a semi-structured interview technique that enabled us to elicit specific rationale for delays in follow-up actions. The interviews were audio-recorded and transcribed. Using an 8-dimensional socio-technical framework, a team of three multidisciplinary researchers with backgrounds in medicine, pharmacy and psychology coded the transcripts using qualitative content analysis. The researchers met multiple times to seek group consensus and codes were refined based on discussions. We used a deductive approach to reduce the data to substantively relevant categories.

Results:
We found 7 dimensions of the socio-technical framework (all except “systems measurement and monitoring) to be relevant for ensuring follow-up of abnormal test results. The technology-related dimensions (hardware and software, clinical content and user interface) included specific factors that could contribute to lack of timely follow-up, such as inability to share information across the systems, inability for patients to access electronic results, lack of prioritization of alerts by the EHRs, lack of multiple patient views in the display, lack of functionalities for task coordination among team members and lack of user-centered design in the EHR. The dimensions related to internal and external environment (internal organizational policies, procedures and culture and external environment) included factors such as inflexible internal organizational policies for test results follow-up, difficulties in delegating responsibilities within the organization and difficulty in external care coordination with outside providers. Clinical staff shortage (“personnel” dimension) was also seen to be a contributing factor. Moreover, we found that overall the rationale for delays in follow-up was more significantly related to the “clinician’s workflow” dimension. Factors such as excessive ordering of unnecessary tests, time constraints to deal with delayed results, difficulty in notifying patients, unique patient’s conditions requiring different prioritizations (such as very sick patients requiring more attention) and lack of standardized workflow processes (such as personalized work arounds to deal with follow-up) contributed to missed abnormal test results follow-up due to variations in clinician’s workflow.

Conclusion:
A broad range of socio-technical factors may lead to delays in abnormal test result follow-up. Enhancing the design of current EHRs and focusing on a host of sociotechnical factors that we identified may improve follow-up of abnormal test results in outpatient settings. In order to improve patient safety, more research is needed to understand the underlying causes for delays in abnormal test results follow-up.
Informing the Requirements for IT-Based Fall Risk Management Solutions in Home Health Care

Onimi Jademi, MS, Uchenna Uchidiuno, MS, Dari Alhuwail, MSc, and G"unes ¸ Koru, PhD
University of Maryland, Baltimore County, Baltimore, Maryland, USA

Significance: Emergency care for injuries caused by a fall is the most frequently encountered potentially avoidable event in home care. Falls can easily lead to major health problems or death. Therefore, managing fall risks becomes an integral component of quality care delivered by home health agencies (HHAs). Currently, most agencies do not adopt or customize IT solutions to specifically address fall-risk management. Among those which do, a thorough understanding of how performance can be improved in fall-risk management is not clearly understood before IT adoption.

Objective: This study aimed at informing the requirements for IT-based fall-risk management solutions by uncovering the challenges and opportunities of achieving better fall-risk management in home care.

Methods: A secondary document analysis of fall and near-fall incidents was performed by examining the incident reports provided by a Medicare-certified not-for-profit HHA located in the mid-Atlantic region of the United States. The reports contained rich and contextual data such as severity of fall, availability of caregivers, patient’s mobility status, and attempted actions prior to fall. The reports were analyzed by two of the researchers using the Framework method. The analysis was guided by the previously established key performance improvement domains to identify the challenges and opportunities to achieve better fall-risk management. A high inter-rater reliability was achieved between the raters: 83% percent agreement and a Cohen’s Kappa of 0.76. The Institutional Review Board approval was obtained before carrying out this study.

Results: The results show that engaging patients and their caregivers, as well as adequately managing their expectations is essential for improved outcomes. Patients had limited understanding of their physical limitations caused by their conditions, and therefore did not comply with the recommendations of using durable medical equipment (DME). Some patients refused to seek help when they needed it. In most cases, caregivers were unavailable; in instances where caregivers were available, they insufficiently monitored patients or were inattentive to their needs. Often, patients and caregivers lacked knowledge of fall-risk management and environmental risks. Some caregivers were unaware of which activities patients were able to perform safely. Some caregivers had their own physical limitations and health problems; therefore they were unable to provide adequate support for the patients. Additionally, poor coordination of interventions, such as delays in the delivery of DME and scheduling lapses, contributed to fall incidents. Other challenges included untimely delivery of care services and incomplete care delivery activities e.g. not following up on medication changes or DME orders.

Conclusion: The results inform the requirements for IT-based fall-risk management solutions, by providing insights into the challenges and opportunities for improving fall-risk management in home care. The HHA could better integrate the EHR and the incident reporting system to facilitate better coordination of care. It was evident from the results that poor engagement was a contributing factor in many fall and near-fall incidents. Leveraging patient portals to deliver relevant information to the patients and their caregivers, can improve their health literacy and increase their levels of engagement in the care plan. Health IT vendors can use this evidence to better understand the needs, and challenges of home care patients and clinicians to develop more effective solutions for fall-risk management.

References
Big-Data Method to Advance Public Health Informatics: Smarter Public Health Prevention System (SPHPS)

Arash Jalali, MPH/PHI, MSHI¹, Douglas Rahden¹, Joerg Heintz¹, Greg Filla², Jason DiNovi², Edward K Mensah, PhD¹, ¹University of Illinois at Chicago School of Public Health, Chicago, IL, United States, ²DePaul University, Chicago, IL, United States

Abstract

The Smarter Public Health Prevention System (SPHPS) provides smarter predictive analytics through big data and cloud computing for reducing health disparities among the working poor and enables smarter workforce health protection or occupational health.

Introduction

This poster presentation will demonstrate public health big-data analytics and integration of health-related information from CMS Virtual Research Data Center (VRDC), Occupational Safety and Health Administration (OSHA), National Emergency Medical Services Information System (NEMSIS), Center for Employee Health Studies Workers’ Compensation (CEHS-W.C.) and external search engine data including Google Health Trends.

Figure 1. Infographics representation of Smarter Public Health Prevention System (SPHPS).

SPHPS¹ is based on the theory that the gravitational force between two objects, which are population health and patient-centered care, is workforce health protection or occupational health. SPHPS integrates CMS VRDC, OSHA, NEMSIS, CEHS-W.C., Google Health Trends and IBM Watson Analytics to improve the health status of low wage workers and health of the general population.

References

Does Health Status Affect Patient Preferences for Sharing Clinical Data for Research?

Imho Jang¹, Diana Guijarro², Jimmy Quach¹, Jihoon Kim, MS², Hyoeun Kim, RN, PhD², Elizabeth Bell, MPH², Robert El-Kareh, MD, MS, MPH², Lucila Ohno-Machado, MD, MBA, PhD²

¹Division of Biological Sciences, UC San Diego, La Jolla, CA, ²Health System Department of Biomedical Informatics, UC San Diego, La Jolla, CA, ³Department of Bioengineering, UC San Diego, La Jolla, CA

Introduction and Background

To improve current practice of obtaining informed consent from patients, we implemented iCONCUR (informed CONsent for Clinical data and biosample Use for Research), a prototype of web-based tiered informed consent system, and surveyed 126 patients recruited from two outpatient clinics – an HIV clinic and Internal Medicine (IM) Clinic – at UC San Diego Medical Center to determine their data sharing preferences. This survey showed that the patients from HIV clinic were more inclined to share their medical data than those from IM clinic. Also, the patients who reported their perceived health status as poor or excellent tended to be more willing to share their medical data for research than those with perceived health status of fair, good, or very good. To better understand the impact of health status on patient’s data sharing decision, we investigated whether there is an association between health status objectively measured with Charlson Comorbidity Index (CCI)¹ and data sharing preference.

Methods

We calculated CCI scores of 126 patients who indicated data sharing preferences in iCONCUR based on their visit diagnoses and medical histories. We also explored the association between data sharing decision and Health Impact Index (HII) scores, which was claimed to be more sensitive to self-reported health status². In both indexes, higher scores indicate worse health status. The data sharing preferences were represented with data sharing scores that range from 0 to 3, where 0 means no intent to share data at all, 1 means sharing only with researchers from UCSD affiliated medical centers, 2 means sharing with researchers from UCSD affiliated medical centers and from non-profit institutions, and 3 means sharing all data with researchers from any institution (including for-profit). We performed correlation analyses between the data sharing scores and each type of health status scores.

Results and Discussion

We observed a weak but statistically significant correlation between data sharing scores and CCI scores (corr=0.24, p=0.007). No significant correlation was observed between data sharing scores and HII scores (corr=0.05, p=0.61). These results indicate that patients with more comorbidities may be more inclined to share their medical data for research. This is consistent with the high level of willingness for data sharing observed among the groups of patients with perceived poor health status. However, it does not explain the high level of willingness for data sharing observed among those with perceived excellent health status. Of note, neither of the health status indexes showed a statistically significant correlation with self-reported health status. This study was done in small scale thus findings are not conclusive. Nonetheless, the results of this study suggest that data sharing decision is a rather complex process influenced by one’s health status, but also by other factors. We have recently launched a larger scale of iCONCUR Phase II study, which will include a larger number of subjects with varying health conditions and sophisticated measures of perceived health status and CCI scores.

Acknowledgement: this study was supported in part by the grant R01HG008802-01(NIH/NHGRI)

Reference

Using Familiar Concepts to Elicit Technology Design Insights

Alvin D. Jeffery, PhD(c), RN, Lorraine C. Mion, PhD, RN, Laurie L. Novak, PhD

1TN Valley Healthcare System, U.S. Department of Veterans Affairs, Nashville, TN, USA
2School of Nursing, Vanderbilt University, Nashville, TN, USA
3Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Introduction

Predictive analytics are gaining popularity in clinical settings and are being used to identify factors related to high-cost patients, readmissions, triage, acute decompensation, adverse events, and treatment optimization. Although statistical outputs of these predictive analytic models can be highly accurate, nurses’ perceptions and information display preferences of prediction/probability-based information are relatively unknown. Our aim was to explore nurses’ perceptions of predicting outcomes (within a larger study on clinical deterioration) by investigating how nurses respond to probability-based information in more familiar scenarios, i.e., using precipitation probabilities from weather predictions.

Method

We conducted 15 semi structured individual interviews and 1 focus group with direct care provider nurses (n=10) and charge nurses (n=8). Each interview lasted 60 minutes and was conducted by 1-2 interviewers. The focus group lasted 2 hours and was conducted by 2 moderators. Explored themes included: (a) perceptions of risk, uncertainty, and probability of adverse events and (b) participant interpretations of weather prediction scenarios. One of the philosophical foundations of qualitative research is that participants, in order to provide valid results, have previously experienced the phenomenon of interest. Because participants had prior exposure to probability-based weather information, we used this as a surrogate to understand how they might use and interpret probability-based decision support tools in the future. Thus, in addition to asking for participant definitions and understandings of risk, uncertainty, and probability, we asked them to interpret 4 different weather scenarios that were consistent across respondents.

Scenarios and accompanying images (see Figure for example) were created to elicit responses from participants that we expected nurses might use when inferring a patient’s likelihood of experiencing clinical deterioration from a predictive tool. An example question was: “Given the following weather scenario, what are your thoughts on planning a major outdoor event?”

Results

Responses to weather scenarios revealed: (a) the desire to review trends rather than absolute values in isolation, (b) an interest in knowing the source of information and type of variables used to develop predictions, and (c) a preference for thresholds to recommend action. When asked about the use of clinical prediction models developed from real-time analysis of electronic health record (EHR) data, almost all participants expressed interest in having access to the predictive information. Participants expressed heterogeneous responses with respect to expression of risk, uncertainty, and probability with some unable to discriminate between the three terms and others providing clear definitions and examples. We hypothesize that interpretation and use of these words are largely acquired from the words their professional peers use rather than a concept about which an individual cognitively processes.

Discussion and Conclusion

By better understanding nurses’ current workflows and perceptions of risk, we hope to achieve greater success in developing novel, probability-based clinical decision support tools. Understanding how nurses cognitively process risk data can contribute to improved design and deployment of predictive tools in the clinical setting and informs important expansions of nursing educational curricula. Given respondents’ positive attitudes toward the proposed availability of clinical prediction models, we plan to examine nurses’ use of prediction models in simulated and real-world settings in the near future.
Utility of the Fast Healthcare Interoperability Resources (FHIR) Standard for Representation of Non-Query Data Mappings in the Arden Syntax

Robert A. Jenders, MD, MS, FACP, FACMI
Charles Drew University & University of California, Los Angeles, CA

Abstract

Context: Arden Syntax encodes knowledge as Medical Logic Modules (MLMs) but lacks a standard data model. Objective: Assess the Fast Healthcare Interoperability Resources (FHIR) standard, previously shown to represent MLM query data elements, for representing non-query data mappings. Methods: 325 MLMs containing 1457 non-query mappings were examined. Result: FHIR can be used to represent all of these data mappings. Conclusion: FHIR adequately represents non-query data mappings in a robust corpus of Arden Syntax MLMs.

Introduction

Arden Syntax is an American National Standards Institute (ANSI) formalism supervised by Health Level Seven International (HL7) for representation of procedural medical knowledge with the goal of facilitating sharing units of knowledge known as MLMs. Some site-specific changes must occur in order for a knowledge base to be transferred from one site to another. Key to minimizing site-specific changes is the standardization of mappings to local clinical data. While these mappings are primarily queries to retrieve data for processing in an MLM, they also include non-query data mappings such as event statements that are used to define MLM triggers and destination statements that route MLM output. HL7 FHIR is a new framework that allows references to data to be defined, represented in XML and bound to terminologies in modular components known as “resources.” For example, concepts such as “patient,” “medication” and “observation” are key FHIR resources, each with structured attributes that may be other resources. Prior work has demonstrated that FHIR is adequate to represent query data elements in a sample of MLMs.

Methods

A robust convenience sample of MLMs was examined. The non-query data mappings were extracted from these MLMs, and the data elements therein were identified. Each then was assessed whether it could be represented—that is, whether the mapping data element could be assigned to a FHIR resource attribute—using FHIR version 1.0.2.

Results

A total of 325 MLMs were pooled from 5 source decision support systems. MLMs concerned mainly with lab tests were the most common (137/325 = 42%), followed by clinical assessment or classification (67/325 = 21%) and medication (41/325 = 13%). The remainder addressed administrative and miscellaneous topics. A total of 1457 explicit non-query data mappings were identified – 430 event statements (30% of all mappings) and 1027 destination statements (70%). The data elements therein were compared to current FHIR resources to assess whether they could be represented. The most common event mappings were hospital admission (representable by the Encounter resource), laboratory test result (DiagnosticReport resource) and ordering of a medication (MedicationOrder resource). While many MLMs had an unmapped, implicit output destination (usually the patient’s electronic chart), the plurality (429/1027 = 42%) of the explicit destination statements were email (telecom attribute of the Person resource), while most of the rest were special project logs, system queues and database tables. The latter could be represented by the Location resource if the extensible value set for Location.type were extended.

Conclusions

FHIR is adequate to represent the data elements found in a large set of non-query data mappings in a robust corpus of Arden Syntax MLMs, although it could be improved to represent specialized system storage locations better. Consideration should be given for use of FHIR to represent data elements in a standard way in the Arden Syntax in order to facilitate procedural knowledge sharing for clinical decision support.

References

An NLP Extension to the Quality Data Model for EHR-Driven Phenotype Algorithm Authoring and Execution

Guoqian Jiang¹, MD, PhD, William K. Thompson², PhD, Luke V. Rasmussen², Richard C. Kiefer¹, Jennifer A. Pacheco², Huan Mo³, MD, Peter Speltz³, Joshua C. Denny³, MD, MS, Jyotishman Pathak⁴, PhD
¹Mayo Clinic College of Medicine, Rochester, MN; ²Northwestern University, Chicago, IL; ³Vanderbilt University, Nashville, TN; ⁴Cornell University, New York City, NY

Introduction
Unstructured data in electronic health records (EHR) has been increasingly recognized as an important source for enabling accurate phenotyping [1-2]. Natural language processing (NLP) tools have been widely used for the purpose of phenotype identification and extraction from clinical narratives. Previous studies [1-2] have demonstrated that many phenotype algorithms include text queries (e.g., keywords, concepts, or regular expressions) to facilitate pattern matching of clinical narratives to be used in conjunction with NLP engines. The Quality Data Model (QDM) [3] has been adopted by the PhEMA project [4] as an information model to support standard representation of phenotype algorithms. However, QDM lacks features in support of representing the NLP components of phenotype algorithms. The objective of this study is to develop and implement an NLP extension to QDM to better support EHR-driven phenotype algorithm authoring and execution.

Methods and Results
We followed the QDM design principles with the notions of category, datatype, attribute and value set, and defined a QDM extension based on collaborative feedback from phenotyping researchers and a literature review [5]. As an initial version, the extension specifies one single category “NLP Extension”, one datatype “Unstructured Data, Document” and 7 attributes (see Table 1). It is expected that most attributes may be associated with a value set with enumerated values or using standard vocabularies. For example, the permissible values in the value set associated with the attribute Document Section can be defined using the HL7 Clinical Document Architecture (CDA)/SNOMED CT/LOINC section type or the section type used in any local EHR system. The data elements of the NLP extension have been loaded in our data element repository and implemented in the PhEMA authoring and execution applications. We tested the utility of the NLP extension by successfully creating a QDM representation of a NLP component in a real-word Benign Prostatic Hyperplasia (BPH) phenotype algorithm developed within the eMERGE consortium, i.e., NOT: “Unstructured Data, Document: Keywords for cancer patients” (Document Section: Problem List). This is an exclusion criterion that describes the exclusion of cancer patients using keywords of “prostate cancer, malignant tumor of prostate, bladder CA, and bladder cancer” in Problem List. In summary, we designed an NLP extension of QDM and implemented it in our PhEMA applications. We plan to collaborate with clinical research informatics and standards communities to rigorously examine the utility of this proposed extension to QDM.

Acknowledgement: This work was supported in part by funding from R01 GM105688 and U01 CA180940.

References
An Evaluation of Activity Trackers for Monitoring Parkinson’s Disease Patient Outcomes
Josette F. Jones, PhD¹, Hwanmei Wu, PhD¹, Jay Patel, BDS,MS¹², Suranga N. Kasthurirathne, BEng¹, Nicole Thai, BS¹, Sunanda Mukherjee, Btch¹
¹Indiana University-Purdue University School of Informatics and Computing, Indianapolis, IN
²Indiana University School of Dentistry, Indianapolis, IN

Problem: Parkinson's disease (PD) is the second most common neurodegenerative disease in the United States. PD results in adverse outcomes including motor impairments as well as non-motor impairments affecting cognition and sleep. Since PD is a neurological disorder, medication has a limited impact on treating PD. Previous research showed that exercise could be beneficial to decrease disease progression and reverse the onset of PD symptoms. However, there has been a little systematic effort to identify the impact of different exercise methods, intensity, frequency, and duration in enabling better PD patient outcomes.

Purpose: The long-term goal of this study is to demonstrate the impact of the vigorous and intensive activity on the progression or recuperation of PD. For this purpose, we propose a pilot study to discover the activity tracker (a) most suitable for use by PD patients and (b) can accurately recognize patterns for different types of activities.

Methods: Activity tracker devices widely available in the market were used to monitor PD patient exercise behavior. We performed literature searches to identify information such as cost, longevity and major functionality of existing activity trackers and used these to compare them against one another. The top devices will be selected and applied to track PD patient exercises. The PD patients will be divided into two groups: the control group with no PD and the intervention PD patient group. The devices will be used to track their activities they engage in, as well as the progression of their PD symptoms.

Results: We present the results of our initial investigation. We evaluated a number of devices including Fitbit, Jawbone, and MotionWatch on criteria including HIPAA compliance, PC software compatibility, the activities/tasks they could measure and many others. Based on this analysis, we identified five devices most suitable for further evaluation. These devices will be tested on patients recruited from RockSteady Boxing (RSB), and used to evaluate how exercise could improve PD outcomes.

Discussion: We found that many activity trackers available in the market were unsuitable for use due to wearability issues/restrictions for PD patients. Additionally, many trackers did not provide direct method to access sensor data that would be used to identify patient activities. However, off the shelf trackers were capable of being used for tracking patient activities, and could be used to monitor patient health. We anticipate that our findings could inform the future design of activity trackers more suitable for use in healthcare management.

Reference:
Assessment of radiation oncology clinician information needs in radiotherapy treatment planning

Alan M. Kalet, PhD, Rebecca J. Hazen, Mark H. Phillips, PhD
University of Washington, Seattle, WA

Purpose and Introduction
The purpose of this study was to characterize the needs of radiation oncology physicians and medical physicists. Radiation oncology is a complex, innovative, and technology driven medical domain utilizing several interconnected software and hardware systems to deliver high energy radiation beams to cure and/or palliate cancer patients. In this study, we sought to glean insights and development directions for informatics tools to improve clinician access and understanding through presentation of historical data and other critical sources of information.

Methodology
45 physicists and physicians working in the Department of Radiation Oncology at the University of Washington Medical Center were solicited to complete a 21 question web survey, providing both structured and unstructured feedback. We asked about the types and sources of information necessary, and clinicians’ perceived ability to obtain that relevant treatment information in planning and decision making activities.

Results
We found that all clinicians (17 total responding – 37.7% response rate) make use of historical patient data when making decisions on current patient treatment plans. Moreover, we found that physicists report better access to data than physicians, and that, among physicians, there was a desire for multiple types of treatment plan data conglomerated from varied sources. A comparison of clinical information types desired/used by the two groups studied is shown in Figure 1. Here, we see that physicians prefer information such as Histology and Karnofsky Score (biological factors) more so than physicists, who report utilizing Orientation/Setup Device (technical factors) more.

Conclusions
The results point to distinct information needs among practitioners, and suggests two possible directions for informatics tool development: one application focusing more on technical planning components, and the other focusing on biological components, with both allowing for comparison of a current case in question to sets of historical cases available to the practice. Our results encourage further development of informatics tools in these two areas to meet the needs of radiation oncology clinicians seeking decision-making information.
Socioeconomic and Clinical Correlates of Computer Use in Cancer Patients
Youjeong Kang¹, PhD, Salimah H. Meghani², PhD,
¹School of Nursing, University of Washington, Seattle, WA
²School of Nursing, University of Pennsylvania, Philadelphia, PA

Abstract: Computer web-based interventions have been increasingly employed. However, they have not been fully effective. Little is known which underlying barriers may affect successful utilization of technology in research setting. This study is a secondary data analysis to describe clinical and sociodemographic correlates of difficulty with computer use in patients with cancer. The findings of this study show that those who are comfortable using computers are wealthier, more educated (i.e. have college degree), are frequent internet users, and have least pain of 3 or less in past week. Future study is warranted to assess whether pain least score is a predictor of comfort even among those who were wealthy and educated.

Introduction: Computer web-based intervention studies across different disciplinary has been increasing. Also, online surveys have become a common method to evaluate the effectiveness of interventions. In particular, as smartphone users are increasing, there are numerous health apps available to monitor health-related information such as diet, physical activity, and to recognize any symptoms. Importantly, despite a continuously rapid increase in health apps use, they have not been fully effective. Little is known about socioeconomic (SES) and clinical barriers to of successful utilization of computer web-based intervention for research purposes in patients with cancer.

Objective: To describe clinical and sociodemographic correlates of difficulty with computer use in cancer patients.

Methods: This study is a secondary analysis of a 3-month observational study (N=241) that used a computer-assisted Choice-based-Conjoint experiment to assess disparities in the outcome of cancer-related pain treatment. Although having internet access was not one of the exclusion criteria, there was also a question about asking internet access at home (N=195). There were four computer use-related questions; difficulty levels of completing the survey using a computer, comfort levels of using a computer, frequency of internet or e-mail use, and preferred method of completing survey (computer vs paper and pencil). Descriptive statistics and bivariate analyses (Fisher Exact/Chi-squared) were conducted in SAS 9.4.

Results: The mean age of the sample was 53 ±11. More participants aged 60 and younger (younger adults) had internet access than those aged older than 60 (older adults). Female participants were 54% of the sample. More White participants had internet access than African Americans (34%). Nearly half of participants had a college degree, and 32% had less than a high school degree. Nearly half of the participants with internet access reported greater than $50,000 of income, and they were the most frequent internet users (p<0.00). Most younger adults (81%) reported that they were comfortable using a computer (p<0.00), and tended to use internet more frequently than older adults (p=0.03). Also, participants with “pain least” scores of 3 or less out of 10 in the last week (lower scores indicate greater pain relief) reported that they were comfortable using a computer (p=0.03). In terms of preference for completing the survey, participants who preferred using a computer (rather than a paper pencil) reported that it was not difficult for them to complete the survey with a computer (p=0.03); they were comfortable using a computer (p<0.00); they used internet more frequently than others (p<0.00).

Conclusions: The findings of this study suggests that those who are comfortable using computers are wealthier, more educated (i.e. have college degree), are frequent internet users, and have least pain of 3 or less in past week. These may inform future computer web-based research about screening profiles (e.g., SES and pain levels in past week) to determine which subgroups of cancer patients who would be more likely to use computer and those needing assistance with research participation in studies that leverage technology. Future study is warranted to assess whether pain least score is a predictor of comfort even among those who were wealthy and educated.

References
Self-management apps for breast cancer survivors: Applying clinical terminology standards for personal decision support

Akshat Kapoor, Ph.D.$^{1}$, Priya Nambisan, Ph.D.$^{2}$

$^{1}$East Carolina University, Greenville, NC, USA
$^{2}$University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Introduction

Most modern electronic medical record (EMR) systems today utilize advanced terminology or coding standards, such as SNOMED-CT, ICD, etc. These standards allow data sharing, accurate coding and billing and overall efficiency in various e-Health functions. By implementing a standard terminology (SNOMED-CT) in a web app for breast cancer survivors, we demonstrate how it may be utilized to provide personal decision support.

Purpose

The system aims to integrate disparate breast cancer survivorship care plans for use in an online web app, by utilizing standardized terminology (SNOMED-CT), in order to provide personalized alerts and reminders for breast cancer survivors.

Problem addressed

Upon completing treatment, a patient is usually provided with a breast cancer survivorship care plan by their provider, to aid in self-management and follow-up care. However, various providers may utilize different terms to describe the same procedure, treatment or symptom in their respective care plans. There is a need to integrate the diverse vocabularies in survivorship care plans, by mapping them to a single concept in a standardized terminology, to allow for the implementation of rules for decision support.

Method

We developed a web app for breast cancer survivors, called ACESO (After Cancer Education and Support Operations) to aid patients in self-management of treatment related symptoms. Rules for alerts [1] were created based on clinical guidelines issued by the ASCO. We utilized a sample of six breast cancer survivorship care plans available online and manually identified each term that described a procedure, test, treatment or symptom. Each term was then mapped to a single concept unique identifier (CUI) in SNOMED-CT (Figure 1). System’s performance in generating the alerts was tested using a randomly generated set of 100 patient scenarios.

Results

From preliminary testing, we were able to identify 52 terms in a set of six different survivorship care plans, which were mapped to 37 unique concepts in SNOMED-CT. The system’s accuracy, recall and precision rates were 90%, 84% and 47% respectively.

Conclusion

This system demonstrates the current variations how incorporating a standardized terminology paves the way for further value-added features, such as personal decision support in the domain of consumer health apps. Further research is required on how these systems can be made more accurate, accessible and efficient for wider use.

References

Abstract
Mild traumatic brain injury is a prevalent and sometimes chronic condition in military service members. Currently, there is a gap in our understanding of the short-term patterns of health care utilization following traumatic brain injury. We focus on injury-related health care encounters recorded in the military electronic health record. Analysis of utilization clusters against patient demographics provides a comprehensive summary of short-term changes in health care utilization following a first mild traumatic brain injury, which may inform us of variation in health outcomes and provide an estimate of resource consumption.

Introduction: A mild traumatic brain injury (mTBI) or a concussion is a poorly understood injury that can alter the way the brain functions. During the last decade a significant amount of attention has been given to the acquisition of clinical data from patients that have suffered mTBI and psychological health (PH) problems after a concussion. Despite widespread prevalence of mTBI in deployed and non-deployed military service members, little is known about health care utilization post-injury and how the frequency and characteristics of the utilization can be used as a measure of outcomes. Through analysis of existing electronic health record medical encounters, we determine relationships between demographic characteristics and patterns of health care utilization by military service members.

Background: Approximately 327,000 military service members sustained a traumatic brain injury (mTBI) between 2000-2015, an estimated 15-20% of who will experience chronic sequelae, resulting in increased health care utilization. The modern active duty military is younger, has a higher percentage of females, and deploys with greater frequency than service members from previous wars. Few studies of the relationship between mTBI symptomatology and health care utilization in the active duty population exist, and the impact of chronic mTBI symptomatology on health care utilization within the Military Health System is unknown.

Methods: A combination of rule-based filters and data-driven models was used to clean more than 14 million EHR encounters from service members diagnosed with traumatic brain injury, resulting in a final population of 109,107 active duty patients with more than 4 million TBI-related clinical encounters.

Results: Preliminary results indicate substantial increases in utilization of outpatient services within 3, 12, and 24 months of injury, with high levels of utilization related to somatic, affective, and cognitive symptomatology. Ongoing analysis is focused on utilization and comorbidity variation by population clusters, as well as potential relationships between short-term utilization intensity and long-term clinical outcomes.
Data Mapping in Development and Implementation of Clinical Data Management Technology for Clinical Research

Shogo Kato, Ph.D.¹, Yoshihiko Morikawa, Bachelor², Seiji Mitsui, Bachelor¹, Takeshi Kuriyama, Bachelor¹, Takahisa Ogasawara¹, Kazuyuki Saito, Bachelor¹, and Naohisa Yahagi, Ph.D.¹

¹National Center for Clinical Research and Development, Tokyo, Japan; ²Tokyo Metropolitan Children's Medical Center, Tokyo, Japan

Introduction

Advances in information management technology create opportunities for researchers to more easily share information in clinical research. Such technology allows clinical research and patient care to become more integrated and interactive. Data sharing is necessary to improve the quality of healthcare and to accelerate the progress in clinical research from bedside to community.

As the Electric Medical Record (EMR) in Japan has been evolved from ordering system, items in EMR are not standardized, because it has not been designed to share data among multiple sources. In the previous trial, the average number of items with standard code of prescription, injection, laboratory and disease, among nine institutes, was 1796 ± 447(mean ± SD), 782 ± 220, 3801 ± 1738 and 23580 ± 7887, respectively. The percentage of mapped items was 12.0 ± 14.0%, 1.7 ± 2.0%, 2.8 ± 1.4% and 93.4 ± 5.0%, respectively. Therefore, collecting, integrating and reconstructing the high quality data into a centralized system from many institutes is challenging.

This study aims to evaluate the importance and effectiveness of code mapping to assemble data for clinical research from different formats and multiple data sources into a centralized system.

Method

The Division of Data Science for Clinical Research at National Center for Child Health and Development, Tokyo, Japan, implemented clinical data management system (CDMS), which is a user-friendly, versatile, and secure network system for clinical data. The data was actually collected from 35 clinics and 11 hospitals geographically dispersed whole Japan. There was no such report before, at least in Japan.

The CDMS used a metadata-driven model that contained built-in knowledgebase to control medical terminology. It was well-constructed knowledgebase consisted of multi-indexed data such as physical examination, laboratory, diagnosis and treatment. The data dealt in HL-7 standard, however, the code for items from EMR such as drug, ICD and laboratory was various in each institute or vendor.

The items in the central database were mapped into the standard code from local codes, by one engineer in systematic method using a supporting software. In addition, items were mapped in manual processing by the engineer, through communicating with each institute and vendor, to improve and ensure the quality of clinical data. The improvement in the percentage of mapped items among actual registered data in each classification was evaluated.

Results

By the operation, the proportion improved to 94.0 ± 3.0%, 94.5 ± 4.2%, 62.9 ± 21.9% and 93.6 ± 4.9%, respectively. In the actual running data, the percentage was 92.2 ± 6.8%, 93.9 ± 9.1%, 65.8 ± 12.6 and 82.1 ± 38.3%, respectively.

Conclusion

The mapping rate before operation was extremely low and it was dramatically improved by the operation both in systematic and manual processing. Development and implementation of clinical science technologies demand a close interaction between clinical research and clinical practice. The global standardized master code, which is independent from vendors or institute, should be developed to share information more easily.
Design, Implementation, and Maintenance of Relevant ASAP Data Report Tools for Patients Presenting to the Emergency Department

Chris Katsura, MD, Frank C. Day, MD MPH, Lynnell McCullough, MD
David Geffen School of Medicine at UCLA, Los Angeles, CA

Abstract

Few clinical decision support tools exist to extract in real-time the pertinent portions of a patient’s voluminous medical record. This is problematic in the Emergency Department (ED), where clinicians face time constraints reviewing extensive records when evaluating a patient. We have created a modular tool that automatically gathers and displays relevant data in a concise format to improve the quality of patient care in the ED, with the ability to expand into ambulatory settings.

Introduction

Healthcare providers today are often inundated with massive amounts of stored electronic data. This can be particularly problematic in the Emergency Department (ED), where clinicians constantly face constraints on the amount of time that can be allocated to combing through the electronic health record (EHR) comprised of terabytes of data, potentially missing crucial information. At our institution, relevant data is often generically entered as progress notes or stored under various categories, depending on the generating service. Chest pain and fever are frequent reasons for ED visits; fever comprises twenty percent of all ED visits, while chest pain comprises five percent. These patients will typically present with multiple comorbidities and extensive medical records. The likelihood that these complaints represent an acutely life-threatening process may be informed by the results of a wide variety of imaging, procedural, and specialist assessment resources. We hypothesize that inserting a relevant data report into physicians’ standard EHR workflow will result in increased awareness of relevant data sources, and that patients may receive more appropriate antibiotics, workups, and dispositions. For practical purposes, we decided to focus our efforts on patients presenting to the Emergency Department complaining of chest pain or fever, with the ultimate goal of expanding this concept to other medical complaints. The report toolbar was initially trialed in the ED and has since been implemented in various ambulatory clinics and infectious disease clinics.

Methods

A multidisciplinary task force comprised of physicians, EHR vendor technical support, and EHR system builders was formed to develop and evaluate a new clinical decision support tool, in the form of an expandable side bar that automatically gathers relevant patient data from multiple sources and displays it all in one convenient location as a single report. For patients presenting to the ED primarily for chest pain, the most recent cardiac tests (stress, echo, cardiac catheterization, troponin) are readily available with the click of a button. For patients presenting with a fever, the most recent WBC, ANC, microbiology culture and sensitivity, medical history, and filtered outpatient immunosuppressant medications are compiled and displayed in an easily viewable tabular format. Print groups are modified to recall and display desired information in an HTML report. The patient’s medical record is filtered by the following categories: date, exam type, pharmaceutical subclass, and laboratory type. The toolbar has been integrated into the ED trackboard, inpatient patient list, and clinic tabs as an informational report. The toolbar will be shared with other departments to allow further integration into ambulatory clinics and other settings. To determine the effect of the relevant data report tool, physicians will be surveyed on their use of the toolbar in their workflow and the effect on their efficiency.

Conclusion

As a result, the clinician will have a powerful tool to extrapolate relevant information quickly to more accurately treat and risk stratify a patient complaining of chest pain or fever. Furthermore, this tool will streamline overall workflow, boost efficiency, improve the coordination of patient care, improve antibiotic selection in sepsis, and minimize unnecessary cardiac testing. In the future, this model will be implemented across other medical complaints and other medical specialties as an invaluable clinical decision support tool.
Grammar frequency and simplification: when intuition fails

David Kauchak, PhD¹, Gondy Leroy, PhD² and Melissa Just, EdD³
¹Pomona College, CA; ²University of Arizona, AZ; ³Rutgers, NJ

Abstract

We investigate whether a medical writer can simplify text by only changing the grammatical structure. Based on a user study, we find that while the sentences look simpler after simplification, they are not easier to understand. For grammatical simplification, better tools are needed to provide more concrete guidance and feedback.

Introduction

Providing text to patients and health information consumers that facilitates comprehension helps create a health-literate patient group. Over the last decades, readability formulas have been touted as writing support tools, but evidence shows they are inefficient and ineffective [1]. We are systematically examining different text features for their potential for simplifying text. We measure the prevalence of each feature, their relationship to text difficulty and how they can be used to simplify text. In previous work, we demonstrated strong results with term frequency, noun phrase complexity, and grammar frequency. In this paper, we examine grammatical simplification.

Methods

To evaluate sentence difficulty based on grammatical structure, we parsed all sentences in English Wikipedia and counted the frequency of the 3rd level in the parse tree (which we denote the grammar frequency). In earlier work, we found that grammar frequency is indicative of sentence difficulty, even when controlling for other variables. We randomly selected 220 sentences from 11 grammar frequency bins (10 per bin) representing increasingly difficult grammatical structures. The writer was told to simplify each sentence by changing only the grammatical structure, i.e. not changing words to simpler variants. We evaluated the simplicity of the sentences before and after simplification with a user study measuring two metrics: perceived difficulty, measured on a 5-point Likert scale, and actual difficulty, measured using a multiple choice Cloze test over four blanked nouns in the sentences. Each sentence was evaluated by 30 participants on Amazon’s Mechanical Turk.

Results

Over half (52.7%) change to a more frequent frequency bin after simplification and 22.3% stayed the same. The expert writer was able to transform sentences into more frequent structures. The simplified sentences also appeared easier; perceived difficulty increased from 2.13 to 2.35. However, the sentences were not any easier to understand and the Cloze score did not change. Figure 1 shows the scores aggregated by bin. We conclude that writers need more direction to simplify text and are testing similarity functions to guide writers towards simpler structures.

Figure 1. Actual difficulty scores (left, multiple choice Cloze test) and perceived difficulty (right, 5-point Likert) for the original sentences, “Original”, and medical writer simplified, “Simplified”.

Acknowledgements

Research reported in this publication was supported by the National Institutes of Health (NIH) #R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Introduction

Phenome-wide association scans, or PheWAS, entail scanning a set of phenotypes to determine whether any are significantly associated with a genotype of interest\(^1\). PheWAS inverts the methodology of genome wide association studies, which find genotypes associated with a specified phenotype\(^2\). This type of study holds tremendous potential in the current era of personalized medicine – the ability to tell from a patient’s genetic makeup what diseases they are most at risk for will be an integral part of healthcare going forward. As part of a set of teaching modules intended to help clinicians learn to use analytics applications, we created a module that allows the user to interact with a simplified version of an existing R package and accomplish some basic PheWAS tasks.

Methodology

We publish an R package and a teaching module. The R package builds upon an existing R package called PheWAS produced by a team at Vanderbilt University\(^3\). The PheWAS package has complexity that makes it difficult for novice R users; for example, there are multiple functions with over 10 input parameters, which can be daunting for an inexperienced coder. We created an R package, simplePheWAS, that serves as an abstraction layer on top of the PheWAS package. We applied several software design principles laid out in *Clean Code*\(^4\) such as input parameter reduction, descriptive parameter naming, and data abstraction. These principles motivate an R package that simplifies usage and has the capability to explain its output to a less experienced user.

The teaching module uses the simplePheWAS package to teach novices about PheWAS analysis. We built the learning module in an interactive Jupyter (formerly IPython) Notebook\(^5\). This application allows the user to write, run, and see the results of their code within a web browser. We provide several types of functionality: (1) The module begins with an explanation of PheWAS analysis and its uses, then moves to (2) a walkthrough of basic PheWAS tasks – loading existing data or generating an artificial dataset with signal from one or more phenotypes selected by the user, to the actual PheWAS, to visualization and analysis of results – borrowing some workflow from the example code in the original PheWAS package. (3) We also allow them to do a backwards analysis not included in the original package – now that we have found a phenotype associated with the SNP of interest, how often does this SNP imply the phenotype, and can we infer possible causality? (4) Then, we take the user through a series of exercises – they run the full PheWAS pipeline from scratch and attempt to find which of an enclosed set of datasets carries a signal. Through this interactive teaching module, users should become comfortable using PheWAS to find associations in their own datasets.

Conclusion

We have created simplePheWAS, an R package and learning module that will teach a novice user how to run a PheWAS study. Our code is available at https://github.com/ekawaler/simplePheWAS.

Acknowledgements

This work is supported by NIH grant 1R25EB020389.

References

The Impact of a Portal for Parents of Hospitalized Children on Healthcare Team Workflow

Michelle M. Kelly, MD1,2, Peter Hoonakker, PhD2, Brad D. Ehlenfeldt, B. Robert Steingass3, Crystal M. Curry, RN, MSN, CPN2, Shannon M. Dean, MD1,3
1 Department of Pediatrics, University of Wisconsin School of Medicine and Public Health, Madison, WI; 2 Center for Quality and Productivity Improvement, University of Wisconsin-Madison, Madison, WI; 3 University of Wisconsin Health, Madison, WI

Introduction

Patient portals are intended to engage patients in care by providing access to healthcare information and facilitating communication with the healthcare team (HCT)1. Although patient portals have been used primarily in the outpatient setting, recent literature suggests they may be beneficial to engage hospitalized patients and their families in care2. Before endorsing widespread adoption, research is needed to understand how inpatient patient portals are used and integrated into the workflow of inpatient HCTs. This study aimed to evaluate HCT perspectives pre- and post-implementation of a portal application (Figure 1) on a tablet computer given to parents of hospitalized children.

Figure 1. Inpatient portal home screen (MyChart Bedside, ©2016 Epic Systems Corporation. Used with permission.)

Methods

In this cross-sectional pre-post study, HCT members (nurses, physicians and ancillary staff) completed surveys on their portal perceptions before and 6-months after portal implementation at a tertiary children’s hospital. The portal shows real-time vitals, medications, lab results, schedule, education, and HCT information and provides a way to send staff messages/requests. Variables were described and compared using percentages and chi-square tests.

Results

Surveys were completed by 94 HCT members pre- and 70 post-implementation (response rate 94 and 88%). Pre-implementation, many respondents thought parent use of the inpatient portal would increase HCT workload (52%), parent portal users would know lab results before the HCT (65%) and have too many questions (68%), and the HCT would be too busy to incorporate the portal into their workflow (71%). Compared to pre-implementation, significantly fewer HCT respondents thought portal use increased their workload (10%, p<0.001) post-implementation and even less reported parents knew lab results before the HCT (4%, p<0.001). Only 3% reported spending more time responding to parent portal user questions, concerns or requests.

Conclusion

HCT respondents anticipated a negative impact of portal use on their workflow; however, these concerns were not realized 6-months after portal implementation.

References

**Patient Accessible Electronic Health Records Portal**

Shreya Chakrabarti, Christopher Kemper, Huanmei Wu  
IUPUI, Indianapolis, IN

**Problem:** It is difficult for a doctor to keep track of the medical history of each of his patients. Similarly, it is also challenging for a patient to remember all the details of each doctor office visits and previous medical conditions. With the paperless electronic medical records and fast-paced technology advances, it is crucial and feasible for the medical world to stay up to date with the newer advent of these technologies. The adoption of these technologies makes lives easier for physicians as well as their patients.

**Purpose:** This project is to create a secure portal for real-time access to patient health information, which helps to improve clinical outcomes by allowing patients to take ownership of their health records. In addition, this project will perform additional medical information analytics with timely alerts and reminders to help patient better utilizing their health information.

**Methods:** An integrated medical information management system is built for patient to access and manage their health records. The system combined the traditional relational database SQL databases (Also includes MongoDB and QGIS (a geographical database system)). Friendly user interfaces using Ruby on Rails are designed to have patient looking at his/her own records. The Apache middleware will connect the backend database and front end patient portal. Once a patient understands the information available, they can update their health conditions, monitor health change patterns and take appropriate actions in a timely manner. System generated alerts for abnormal or out of range results can further a patient's knowledge and better healthcare.

**Results:** A prototype of the patient portal and the integrated medical information management system have been developed. De-identified patient records are input as the test bed. The patient portal accesses the Patient Database and keeps a track of each patient medical information like the current and past medications prescribed to the patient, the various diagnosis of the physicians during different visits by the patient. It also records the Vitals of the patient updated by the doctor during the patient’s last visit. In addition, it gives the patient a way to update his/her own vitals, these vitals will also have restrictions on them with respect to values. Different dashboards are designed for patient viewing, updating and managing their records. The figure below lists two sample dashboards: one is for the general information about the patient upon logging in and the other is a visualization for the patient to check on the historical changes of the patient vital records. Database security is enforced for patient privacy.

**Conclusion:**
Patient portals help a patient keep track of his/her health over a period of time and be motivated to maintain their vital chart in a healthy range. The patient portal would also help a physician with patient history in case the patient moves to a different state or country.

**References:**


2) http://www.morehead.unc.edu/revitalise/feb/diabetes.gif
Secondary Use of Health Information Exchange: Bivariate Analysis of Return Visits After Emergency Care

Eugene Kim, MD12, Tina Lowry, MS2, George T Loo, MPA, DrPH2, Bradley D Shy, MD2, Ula Hwang, MD2, Nicholas Genes, MD, PhD2, Lynne D Richardson, MD2, Cindy F Clesca, MA2, Jason S Shapiro, MD2

1Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA
2The Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

Introduction:
Health information exchange (HIE) provides a window into patient behavior across medical systems. Early (within 72 hours) emergency department (ED) returns have been used as a quality metric and other types of revisits (e.g. 30-day readmissions) may currently incur penalties for reimbursement. However, motivations for these early revisits and patients’ choice of location are poorly understood. We hypothesize that this novel secondary use of HIE data may determine differences in specific attributes of: 1) patients without early return visits, 2) patients returning to the same initial “index” ED, and 3) those returning to a different ED. Elucidating these qualities may give insight into health care utilization and ways to improve outcomes.

Methods:
De-identified data from 3/1/09 to 2/28/14 were used from Healthix Inc., an HIE with data from 31 hospitals in the New York metropolitan area. Early ED returns were defined as having the first index visit to one ED and a second visit within 72 hours to the same or another hospital’s ED or inpatient service. As per the Association of Academic Health Centers, an academic program was defined as having an affiliation with a medical school having at least one health professions program and an affiliated teaching hospital. A bivariate analysis of the encounters was performed, analyzing the sex and age of the patients, time of the visit, hospital county in relation to the patients’ home county, and the academic status of the hospital.

Results:
The analysis included 12,621,159 total visits. There were 948,972 (7.5%, p<0.001) early returns made by 563,687 unique patients; of these return visits, 11.35% occurred at a site other than the initial hospital.

ED encounters with a return visit to the same or different hospital were 16% more likely to be by males and 48% less likely to be by patients under 18 years of age when compared to encounters that did not result in a return visit. The return visits after an initial ED encounter were 23% more likely to be in the patients’ county of residence and 3% more likely to occur during the day than encounters without return visits.

Visits with early returns to an institution different from the initial index visit were 32% more likely to be by male patients and 64% more likely to be by those under 18 years of age compared to encounters with early returns to the initial index hospital. The initial visits were 24% more likely to occur during the day, while the return visits were 30% less likely to be during the day in comparison to returns with both visits at the same site. An initial visit to a site housing an emergency medicine residency or being located in the same county as the patient was 63% or 9% less likely to result in a different site return visit when compared to initial visits having returns to the same site, respectively.

Conclusion:
HIE data may be secondarily used to characterize patients after visits to the ED. Patients returning within 72 hours to a different hospital were more likely to be male and younger, with the initial hospital less likely to have an emergency medicine residency program. Patients seeking more definitive care, the local incidence of disease, and the existence of fewer pediatric centers could play a role in these findings. Thus, factors associated with the patient, visit, and hospital itself may have influence on early returns and may not necessarily reflect quality of care. Moreover, early return rate calculations without the use of HIE may be misleading since these may miss more than 11% of early returns. Further study using HIE or similar technologies may suggest motivations behind return visits and potential areas for improvement in care delivery.
Developing A Mobile Application to Improve Diabetic Patients’ Self-Care Behaviors: A Functionality Analysis

Min Soon Kim, PhD1,2, Qing Ye, MS1, Uzma Khan, MD3, Suzanne A. Boren, MHA, PhD1,2
1Informatics Institute; 2Department of Health Management and Informatics; 3Department of Medicine, University of Missouri, Columbia, MO

Abstract
We analyzed functions and sub-functions of 168 diabetes mobile apps from iTunes and Google Play according to the validated American Association of Diabetes Educators (AADE7) Self-Care Behaviors™. A majority of the apps were designed to support Monitoring, Healthy Eating, Taking Medication, and Being Active. Fewer apps focused on Problem Solving, Reducing Risks, and Healthy Coping, which should be incorporated to enhance diabetic patients’ self-care behaviors.

Introduction
Changing patients’ self-care behavior improves control and decreases the risk of long-term disability and complications of diabetes. Mobile apps have the potential to improve patients’ self-care behaviors. However, current literature lacks discussion about the design and ability of diabetes apps to change self-care behaviors of diabetic patients. The purpose of this study was to examine the functionality on self-care behavior of current diabetes apps.

Methods
Usability experts, a diabetes educator, and a clinician utilized a multi-step review process. In an online search of two stores with major market share (iTunes and Google Play), three distinct and broad search terms were used to capture a wide range of diabetes apps: diabetes, blood sugar, and glucose. Apps were excluded if they (1) were not designed for diabetic patients, (2) were not written in English, or (3) only provided access to reference material. The functionality of each app was analyzed and classified according to the validated AADE7 Self-Care Behaviors™ by the American Association of Diabetes Educators. The seven behaviors are Healthy Eating, Being Active, Monitoring, Taking Medication, Problem Solving, Reducing Risks, and Healthy Coping. The functions that support the seven healthcare behaviors were then divided into sub-functions. We conducted a descriptive analysis of the functions and sub-functions.

Results
Out of 1,050 apps retrieved, 168 apps were identified as eligible. The majority of the apps were designed to support behaviors of Monitoring (72%), Healthy Eating (71%), Taking Medication (52%), and Being Active (37%), all of which require tracking patient-entered information. On the other hand, few apps explored behaviors of Problem Solving (20%), Healthy Coping (8%) and Reducing Risks (5%), all of which provide management guidance based on patient-entered data. The functions that support seven behaviors were categorized as 70 sub-functions (Table 1).

Table 1. The three most frequent sub-functions according to the AADE7 Self-Care Behaviors™.

<table>
<thead>
<tr>
<th>AADE7 Self-Care Behaviors</th>
<th>Three Most Frequent Sub-functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy Eating (71%)</td>
<td>count carbohydrates, read food labels, measure each serving</td>
</tr>
<tr>
<td>Being Active (37%)</td>
<td>keep track of activity, check blood sugar levels, knowledge of exercise</td>
</tr>
<tr>
<td>Monitoring (72%)</td>
<td>record blood sugar results, log height, weight, or BMI, log blood pressure or pulse</td>
</tr>
<tr>
<td>Taking Medication (52%)</td>
<td>manage medication list, calculator, reminder</td>
</tr>
<tr>
<td>Problem Solving (20%)</td>
<td>alert or reminder for abnormal data, sharing others, analyze daily issues</td>
</tr>
<tr>
<td>Reducing Risks (5%)</td>
<td>foot care, monitor signs and symptoms, knowledge of reducing risks</td>
</tr>
<tr>
<td>Healthy Coping (8%)</td>
<td>record mood, attend support groups, knowledge of healthy coping</td>
</tr>
</tbody>
</table>

Conclusion
Discussion with an experienced endocrinologist and literature review revealed that Healthy Eating, Problem Solving and Healthy Coping are critical principles for improving self-care behaviors. However, Problem Solving and Healthy Coping principles may be more difficult to address in app design than the other five principles. This may attribute to the skewed app development trend. Future diabetes apps should incorporate functionalities of Healthy Eating, Problem Solving, and Healthy Coping, to better support changing diabetic patients’ self-care behaviors.
Data Topography of a Large Multi-Site Research Network
Jeffrey G. Klann, PhD1,2,3, Vijay Raghavan, MS, MBA1, Doug MacFadden, MS1, Sarah Weiler, PhD1, Kenneth Mandl, MD, MPH1,4, Shawn N. Murphy, MD, PhD1,2,3
1Harvard Medical School, Boston, MA; 2Partners Healthcare, Boston, MA; 3Massachusetts General Hospital, Boston, MA; 4Computational Health Informatics Program, Boston Children’s Hospital Boston, MA

Background
The Scalable Collaborative Infrastructure for a Learning Health System (SCILHS, pronounced “skills”) is a growing network of health centers across the United States, presently covering over 10 million patients at eleven sites. SCILHS is a Clinical Data Research Network (CDRN) in the Patient-Centered Outcomes Research Institute’s PCORnet, a national effort to instantiate a ‘network of networks’ that supports large-scale comparative effectiveness research.

The dizzying variety of data representations and terminologies in clinical systems makes developing an interoperable network difficult. To support interoperability in a distributed network, SCILHS has uniformly adopted the PCORnet Common Data model (CDM). However, the network sites do not tend to have their source data in the terminologies required by the CDM. To that end, SCILHS developed a methodology to allow mapping from source terminologies to the CDM ontology, which has been implemented at eight network sites. The implementation gave us insight into the challenges and potential pitfalls of data interoperability in a real-world scenario. Here, we report the data topography of the network and some lessons learned.

Methods
SCILHS uses Informatics for Integrating Biology and the Bedside (i2b2) as its technical backbone. i2b2 is an open-source clinical data warehousing and analytics platform funded by the National Institutes of Health. It is used at over 100 sites nationwide, including several CDRNs and the National Center for Advancing Translational Sciences (NCATS) Accrual to Clinical Trials (ACT) network. i2b2 supports live distributed queries through the Shared Health Research Informatics Network (SHRINE) platform.

i2b2’s data model is highly adaptable to ingest data from various source systems, and local implementations develop concept hierarchies (called “ontologies”) that provide a window into the imported data. For SCILHS, the data are exposed as a single CDM ontology, and each site used our methodology to map from their local terminologies’ codes to the CDM ontology. We assisted the sites with this process and worked to resolve challenges. At completion, we received a report from each site on which CDM ontology terms are available at each site.

Results
The CDM ontology is divided into many clinical domains, each of which has its own coding system requirements. Major terminologies used are listed below, along with a summary of challenges encountered.

- **ICD-9:** All sites had at least some of their diagnoses and procedures in ICD-9 format (due to billing requirements). However, the problem list sometimes used non-standard codes that required manual mapping.
- **CPT:** Most sites had some procedure codes in CPT format. However, several sites discovered some procedure codes were stored in separate billing systems for professional staff, which were not available to researchers. Also, retired CPT codes were not initially placed in our CDM ontology, making some procedure categories appear empty when an old code was being used in historical data.
- **RxNorm:** Sites did not tend to store their data in RxNorm, but often a crosswalk to RxNorm was available. However, these were frequently incomplete and resulted in some data loss. At some sites, only NDC was available, and even with a variety of terminology resources for mapping NDC to RxNorm, the sheer number of NDC codes make maintaining these resources difficult, leaving NDC to RxNorm mappings incomplete.
- **LOINC:** Many sites involved in research had invested in mapping their laboratory codes to LOINC already. However, frequently the LOINC codes used were inconsistent between sites and did not match the expected codes specified in the CDM. This problem was not caught for some time, because the availability of LOINC codes created a false sense of security surrounding data completeness.

Discussion and Conclusion
A wide variety of subtle terminological and data mapping issues hamper interoperability. SCILHS has gathered significant experience with these issues, and we lay out some highlights above. The largest problems encountered were *historical (retired) codes, inconsistent mappings, non-standard local codes, and data availability.*
Incorporating Televisions into Older Adults’ Health Routines: A Case Study

Laura Kneale¹, MS, Hilaire J. Thompson², PhD, RN, CNRN, ACNP-BC, FAAN, George Demiris¹,², PhD, FACMI
¹Biomedical and Health Informatics, ²Biobehavioral Nursing & Health Systems
University of Washington Seattle, WA 98195

Introduction
Older adults spend three times as many hours watching television than younger adults¹. Digital Interactive Televisions (DITVs) are televisions that use an Internet connection to incorporate non-traditional functions. DITV’s for health and social care have been previously studied but no trials have been conducted with older adults in the United States². The purpose of this study was to evaluate the feasibility of using a DITV with health and wellness functions and a companion caregiver portal with community-dwelling older adults and their informal caregivers.

Materials and Methods
We tested a commercially available DITV that was designed for older adult users. Our team was not involved in the design of this device. The DITV allows users to: a) alert caregivers that help is needed, b) place video-calls, c) view pictures and messages, d) view a calendar, and e) respond to health alerts. The system had a caregiver portal to manage the DITV functions. This study recruited Seattle-area community-dwelling older adults and their informal caregivers for a 30-day trial. The DITV was installed in the older adults’ home. Participants were trained on the system, and asked to complete an in-person, audiotaped interview at the beginning and end of the trial. We thematically coded the interview transcripts to identify benefits and challenges with using the DITV.

Results
Participant Recruitment and Retention: Despite varying recruiting techniques, it was difficult to identify older adults that were both eligible and interested in the study. Potential participants aged 60-70 years old, were interested in the study but were not willing to identify a caregiver, and potential participants over age 70 often refused to participate because installing the new television may disrupt their current viewing routines. We also had difficulty installing the DITV because of space constraints and Internet availability. We completed one trial during the 12-month study.

Case Study Results: One 28-day trial was completed in May 2015 with a 61-year-old Caucasian female and her adult daughter. The interviews with these participants highlight several benefits of using the DITV’s functions. The participants found the reminders helpful in supporting the desired health routines, and the caregiver felt that the system improved the daughter’s understanding of the older adult’s health behaviors. The participants also reported challenges with the system. The older adult discussed challenges with the alerts firing during family viewing times. In addition, the older adult would be unaware that she missed the DITV-based alerts when the TV was off, as missed alerts were only communicated through the caregiver portal. The participants discussed how a DITV could fit in a household with other technologies, and discussed the information that should be shared between different technologies (e.g. alerts), and the features that should be kept separate (e.g. entertainment and email). The caregiver described her use of the portal, and provided recommendations including allowing the older adult to update the DITV content through the television interface.

Discussion
While ubiquitous objects may end up playing a role in health and disease prevention, the practical implementation of this concept carries some challenges. Replacing older adults’ television, even with another television, may be difficult for some older adults. Our findings suggest disrupting an individual’s television routine may make some older adults unwilling to adopt DITVs. Perhaps a standalone set-top device that can be connected to an existing TV, similar to an Apple TV, would have better acceptance among these older adults. In order to identify these challenges, more testing needs to be completed with older adult end users to assess their acceptance of new devices, even if the devices appear to be similar to items already in their homes.

Conclusion
Our case study suggests that there are many potential benefits of using a television to incorporate health routines into older adults’ homes. Future research is needed to better understand the DITV features that older adults’ desire, how these technologies will best fit into existing technology in the home, and how to encourage older adults to adopt these types of technologies.

References
MagnoMath: Mobile Application for identifying Dyscalculia Symptoms related to Magnocellular Reasoning

Greger S. Knudsen, M.Sc⁠¹, Ankica Babic, PhD⁠¹²
¹ Department of Information Science and Media Studies, University of Bergen, Norway
² Department of Biomedical Engineering, Linköping University, Sweden

Abstract

This poster presents a study that measured correlations between learning disabilities in mathematics and magnocellular reasoning using MagnoMath, a software application for smartphones. Unlike most software aids used in diagnosing symptoms of learning disabilities, which rely exclusively on scoring users’ performance, MagnoMath additionally measures the correlation between these symptoms and magnocellular reasoning. To achieve this, a k-Nearest Neighbor classification algorithm was applied. Field expert evaluation has found the application to be appropriate and productive for its intended purpose. Initial user testing has been positive and provided suggestions for further improvements.

Introduction

The majority of software tools for assisting students with learning difficulties focus on teaching mathematical principles. Often there is no further insight given regarding the causes. This study explored the relationship between neurons believed to be responsible for developing learning disabilities, called magnocellular cells, and underperformance in mathematics [1]. By measuring these correlations, one could improve diagnosing and obtain a greater insight into the condition.

Method

The software applies two parameter groups for identifying and measuring correlation. Selected tasks activating reasoning in magnocellular cells are contrast sensitivity and spatial ability. The latter is the process of mentally manipulating three-dimensional figures. Selected tasks for evaluating dyscalculia symptoms are counting, subitizing and arithmetic. Subitizing is the process of recognizing patterns of objects and immediately determining the quantity without counting them one by one. The parameter groups were classified separately by two k-Nearest Neighbor classifiers using data gathered by given evaluation tasks. The key evaluation result is the correlation computation between dyscalculia symptoms and magnocellular reasoning. This computation applies mathematical reduction crucial to identifying possible reasoning dependencies. The evaluation of MagnoMath was two-folded. Ten software and mathematics field experts from the Department of Biomedical Engineering, Linköping University assessed the application’s usability and behavior. In addition, two learning disability field experts coming from the Department of Behavioral Sciences and Learning, Linköping University, and Dyslexia Norway, Oslo, evaluated the application for its validity and accuracy.

Results and discussion

The classification algorithm applied shown accurate and efficient performance allowing users to run several evaluations while getting updated feedback. Using MagnoMath regularly could benefit learning and understanding. Feedback from field experts stated that the application was easy to use, had suitable features and performed well, stably and efficiently during use, thus proving it to be suitable for its intended audience.

References

Selecting Health Insurance Coverage: Can Complexity and Context Be Used to Manipulate Choices?
Malgorzata Kolotylo MScBA and Monica Chiarini Tremblay PhD, Florida International University, Miami, FL

Abstract. Despite its low trust score, the internet remains to be one of the main sources of information for health insurance seekers (Furtado et al. 2016). Many of the uninsured who constitute the primary target market for private health insurance companies are unemployed or low-income social groups who, although may not use PC to shop for insurance, tend to use smartphones or at least cell phones, thus a substantial portion of advertising and reaching the customer occurs via not only websites but also mobile applications and texting. Under the Affordable Care Act, health insurance is mandatory and those uninsured must apply for a minimum level of coverage or face a tax penalty. However, choice of health insurance coverage is difficult, as the consumer may have to judge costs vs. their medical and financial risks: evaluate his/her future use of healthcare services, assess, compare out-of-pocket costs and premiums provided by different plans while still taking into consideration provider features (such as network, drug coverage of potential exclusions), way of working with a doctor and coverage level (bronze, silver, gold and platinum) (Rice, Jacobson, Cubanski and Neumann, 2014). In order to facilitate the choice of an insurance plan Americans were provided with federal- and state-based Health Marketplaces, which constitute online-based platforms that enable comparison and purchase of health plans. Additionally, there are also private non-governmental websites that facilitate health insurance search, such as healthcare.com. Some of those websites may not provide full information (they usually contain a disclaimer). Healthcare.gov is not open for use outside of the open enrollment period unless under Special Enrollment, Medicaid or CHIP. Thus, under certain circumstances, the consumer has no choice but utilize intermediaries. Privately-owned websites utilize well known marketing techniques to influence choice (Xiao and Benbasat, 2011). Many privately-owned websites receive compensation from third-party providers (advertisers), which in turn may influence the way offers concerning insurance companies are presented to the user. We propose a multi-method study concerning how manipulated information influences consumers’ purchase decision. We seek to understand how consumers can be manipulated in their choice in order to warn and educate the public. Future research may build upon our findings to scrutinize the contextual dynamics that enable online vendors deceive consumers.

Methodology: This research will undertake a multi-method approach: Field interviews and observations will be carried out in order to gain deep insights which include: 1) an examination of private health insurance broker websites 2) interviews with health consumers in order to obtain their perspective and their experiences with insurance purchase decisions. Next, we will conduct a controlled experiment. The dependent variable is constituted by purchase intention, the independent variable is manipulation of information (manipulation of content and presentation) and its impact is mediated by affective and cognitive responses of the buyer (based on Xiao and Benbasat, 2011). The relationships are hypothesized to be moderated by: a) task complexity (high vs low) and purchase situation (impulse vs planned) b) measured variables of health status, prior knowledge of insurance and uncertainty avoidance. During the experiment a random sample of consumers will be asked to purchase insurance online. Sample size will be calculated with a power analysis. A survey will be distributed before and after the task to measure the difference in affective and cognitive mechanisms of the respondents. A website resembling online insurance broker will be designed with information manipulations as specified in research design.

Early Research Findings. The poster will explain the design of the experiments and early findings from interview and pilot study.

References:
An Electronic Health Record Phenotyping Algorithm for Identifying Patients with Hepatorenal Syndrome

Jejo Koola1,2, Samuel B. Ho3,4, Michael E. Matheny1,2

1Tennessee Valley Healthcare System VA, Nashville, TN; 2Vanderbilt University, Nashville, TN; 3University of California at San Diego Medical Center; 4Dept. of Veterans Affairs, San Diego, CA

BACKGROUND: Patient phenotype identification serves as an essential step for outcome prediction, survival analysis, and other retrospective and prospective research studies. The Hepatorenal Syndrome (HRS), one of several cirrhosis associated kidney disorders, stands as an archetype of multi-organ failure; the median survival for Type 1 HRS is two weeks. An International Classification of Diseases version 9 (ICD9) code exists for HRS; however, its accuracy is unknown. ICD 9 codes are well known for being poor markers of patient phenotype. Hepatorenal Syndrome is an especially complex phenotype as it requires ruling out other causes of renal failure in patients with cirrhosis, of which there are many.

OBJECTIVE: We sought to identify the accuracy of the ICD 9 code for Hepatorenal Syndrome and develop a phenotyping algorithm based on electronic health record (EHR) data.

METHODS: We analyzed a retrospective cohort of patients hospitalized nationwide in the Department of Veterans Affairs between 2005 and 2013. We selected 504 hospitalized patients with a prevalent or incident ICD 9 diagnosis of cirrhosis and laboratory evidence of acute kidney injury stratified by renal (AKIN Stage I, II, or III) and cirrhosis (MELD score < 20 versus >= 20) severity. We performed manual chart review of the entire inpatient admission including medical notes, labs, and radiology procedures to categorize patients into one of five phenotypes: HRS Type 1, HRS Type 2, HRS Type Indeterminate, Maybe HRS, and Not HRS. For probabilistic phenotyping, we constructed a data matrix with 704 candidate predictor variables consisting of demographic variables, laboratory values, pre-existing and inpatient conditions and procedures based on administrative codes, and outpatient and inpatient medications. We performed penalized logistic regression using a Least Absolute Shrinkage and Selection Operator (LASSO) to select a subset of the candidate predictor variables. The outcome variable of interest was HRS phenotype by chart review (HRS Type 1, Type 2, and Indeterminate Type were lumped into a single category of Yes HRS). We constructed three models based on the “Maybe HRS” outcome: in the first, observations that were annotated as Maybe HRS were given an observation weight of 0.5 in the LASSO outcome prediction, survival analysis, and other retrospective and prospective research studies. The Hepatorenal Syndrome (HRS), one of several cirrhosis associated kidney disorders, stands as an archetype of multi-organ failure; the median survival for Type 1 HRS is two weeks. An International Classification of Diseases version 9 (ICD9) code exists for HRS; however, its accuracy is unknown. ICD 9 codes are well known for being poor markers of patient phenotype. Hepatorenal Syndrome is an especially complex phenotype as it requires ruling out other causes of renal failure in patients with cirrhosis, of which there are many.

RESULTS: Of the 504 patient cohort, 115 and 28 had HRS Type 1 and Type 2 respectively, 21 had HRS Type Indeterminate, 58 were annotated as Maybe HRS, and 282 did not have HRS. Treating the patients with a “Maybe HRS” annotation as positive for HRS, the discharge ICD 9 code had a sensitivity, specificity, positive predictive value, and negative predictive value of 0.58, 0.91, 0.83, and 0.73 respectively. Conversely, treating “Maybe HRS” annotations as negative for HRS, the ICD 9 code had operational parameters of 0.63, 0.85, 0.68, and 0.83 respectively. Of the 704 candidate predictor variables, the LASSO regression model identified 20 that were predictive of a HRS diagnosis. The model had an AUC of 0.92 (95% CI: 0.91 – 0.94). Table 1 lists the variables along with their odds ratios from the unpenalized logistic regression model. Using Youden’s point to define the optimal threshold, the sensitivity, specificity, positive predictive value, and negative predictive value of the EHR phenotyping algorithm were 0.88, 0.83, 0.80, and 0.90. Sensitivity analysis of the “Maybe HRS” annotation showed no significant difference from the “original model” with AUCs for the two additional models of 0.93 (95% CI: 0.91 – 0.95) and 0.91 (95% CI: 0.89 – 0.94) respectively.

DISCUSSION: Hepatorenal Syndrome is a complication of cirrhosis with high morbidity and mortality. As all studies with HRS have been limited to few patients (between 15 and 100), retrospective analyses of large EHR cohorts may prove useful. However, no published literature exists on the accuracy of the ICD 9 code for Hepatorenal Syndrome. In this study, we show that the ICD 9 code has very poor sensitivity most likely because patients with Hepatorenal Syndrome are coded (if at all) with the more general “Acute Kidney Failure.” We further develop a probabilistic EHR phenotyping algorithm with significantly improved sensitivity while maintaining comparable specificity.

Table 1: Odds ratios and confidence intervals for the variables identified by the lasso model. Odds ratios are derived from rerunning the model in an unpenalized general linear model.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Odds Ratio (95% CI)</th>
<th>Variable Name</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Urine Sodium</td>
<td>0.67 [0.45, 0.99]</td>
<td>Average BUN</td>
<td>0.75 [0.31, 1.77]</td>
</tr>
<tr>
<td>Peak Creatinine 48h to Discharge</td>
<td>1.05 [0.62, 1.78]</td>
<td>Minimum BUN</td>
<td>2.89 [1.45, 5.75]</td>
</tr>
<tr>
<td>Creatinine Diff. (1st 48 hours vs rest of adm.)</td>
<td>0.33 [0.20, 0.56]</td>
<td>Average Blood Sugar</td>
<td>1.13 [0.70, 1.83]</td>
</tr>
<tr>
<td>Average Sodium</td>
<td>0.76 [0.55, 1.06]</td>
<td>Average MCHC</td>
<td>0.79 [0.31, 1.96]</td>
</tr>
<tr>
<td>Average Bicarbonate</td>
<td>0.70 [0.49, 1.00]</td>
<td>Maximum MCH</td>
<td>0.89 [0.59, 1.35]</td>
</tr>
<tr>
<td>Average Bilirubin</td>
<td>0.89 [0.61, 1.28]</td>
<td>Minimum Albumin</td>
<td>0.79 [0.54, 1.13]</td>
</tr>
<tr>
<td>Inpatient Midodrine</td>
<td>7.02 [3.06, 16.07]</td>
<td>Minimum INR</td>
<td>1.21 [0.85, 1.72]</td>
</tr>
<tr>
<td># Inpatient ICD 9 codes for sepsis</td>
<td>0.16 [0.05, 0.53]</td>
<td># Inpt. ICD 9 codes -ATN</td>
<td>0.38 [0.25, 0.58]</td>
</tr>
<tr>
<td># Inpatient Charlson-Deyo liver disease</td>
<td>1.36 [0.77, 2.40]</td>
<td># Inpt. ICD 9 codes - ascites</td>
<td>1.37 [0.93, 2.05]</td>
</tr>
<tr>
<td># Inpatient ICD 9 codes for HRS</td>
<td>3.88 [1.78, 8.41]</td>
<td># Inpatient Paracentesis</td>
<td>1.13 [0.82, 1.55]</td>
</tr>
</tbody>
</table>

BUN: Blood Urea Nitrogen; MCHC: Mean Corpuscular Hemoglobin Concentration; MCH: Mean Corpuscular Hemoglobin; INR: International Normalized Ratio; ATN: Acute Tubular Necrosis
Challenges and Triumphs: Developing an NLP Pipeline for Thyroid Ultrasound Reporting

Rajesh Koralkar, MSHI, Eric LaRose, BS, Yiqiang Song, M.S, David F. Schneider, MD, Peggy L. Peissig, PhD

1Marshfield Clinic Research Foundation, Marshfield, WI; 2University of Wisconsin, Dept. of Surgery, Madison, WI

Introduction
Thyroid nodules are extraordinarily common. Approximately 66% of adults will have a thyroid nodule. Thyroid ultrasonography (USG) remains the best method of imaging thyroid nodules and prioritizing them for biopsy, observation, or surgery. Unlike the reporting for other imaging techniques, USG reporting is largely unstructured and variable in pattern. This makes extraction of critical nodule features challenging. We report on the development of a natural language processing (NLP) pipeline to extract thyroid nodule characteristics from thyroid USG reports present in the electronic health record (EHR).

Methods & Results
This study was approved by the Marshfield Clinic Institutional Review Board. The study includes patients who had thyroid USG and Fine needle aspiration cytology (FNAC) within the Marshfield Clinic Health System from January 2007 –September 2015. Patients with indeterminate cytology results were selected for this study. Radiology documents including thyroid USGs, are stored in a structured XML format within the EHR. Descriptive nodule information is embedded within text strings that required NLP. We used NLP to identify terms in radiology reports and to decide their status such as certainty and negation. Study terms were identified in the Unified Medical Language System and extracted using clinical Text Analysis and Knowledge Extraction System (CTAKES), from the processed documents. Similarly, measurements of nodules were identified using customized regular expression operations. The relation between terms and values were then evaluated by their presence in the same sentences to construct logical groupings. (refer to Figure 1: Workflow Overview). Post processor items included thyroid lobe and nodule for extraction of characteristics. We identified 4593 patients who had USG and FNACs done on a thyroid nodule. Of these, 1400 patients had indeterminate cytology results. We manually validated data from 100 documents calculating the percentage of correct extractions for the characteristics for each of study items: measurements (height, width, diameter) 90%; unit of measure 92%; location 83%; calcification 75% and echogenicity 92%. One of the challenges we had is in determining location of the item as most sentences had mention of both Left and Right in context to different items. The missing measurements were attributed to where no study item was associated to measurement in same sentence. CTAKES output for most part Negated terms appropriately.

Discussion
This is an example of a novel NLP application for extracting critical thyroid nodule features from USGs. The unstructured and variable pattern of thyroid USG reporting makes determining calcification status the most challenging. This preliminary work will be extended by improving post-processing of multiple sentence logical groupings and then to another institution to test the generalizability of the NLP pipeline.

References
Analysis and Redesign of Alerts and Reminders in a Commercial Electronic Health Record System to Reduce Alert Fatigue

Polina Kukhareva, MS, MPH, Damian Borbolla, MD, MS, Kensaku Kawamoto, MD, PhD, MHS
Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA

Abstract
“Alert fatigue” is an important barrier to clinical quality improvement through the use of information technology. At the University of Utah Health Care, the Clinical Decision Support (CDS) committee is employing a pragmatic approach to monitoring and managing CDS interventions in the context of a commercial electronic health record (EHR) system. Through the analysis of the top volume CDS alerts and reminders and their associated process and outcomes measures, we are aiming to reduce alert and reminder volume and associated fatigue. Here, we describe our approach to alert management in a commercial EHR system and discuss planned next steps.

Introduction
Many CDS alerts and reminders are inaccurate or contextually irrelevant, causing “alert fatigue” that can lead to ignoring appropriate CDS interventions, threatening patient safety.[1] While recognized as an important problem, many healthcare IT departments are unable to dedicate significant resources to address this issue, due to competing demands for meeting regulatory requirements or simply meeting new end-user requests. Therefore, there is a need for a pragmatic approach to address alert fatigue problems. Here, we share our experiences in identifying and modifying low accuracy and high volume alerts and reminders. To improve the specificity of our alerts and reminders and to reduce alert fatigue, we (1) analyzed success level of the alerts and reminders and (2) proposed pragmatic methods to improve specificity.

Methods
University of Utah is currently using a commercial EHR system (Epic). Medication alerts are supported by the Medi-Span drug database. Information about alerts history is recorded in the EHR and transformed into the data warehouse. We performed a descriptive analysis of all medication alerts and other custom alerts and reminders (known as best practice advisories or BPAs) which fired during 2015. For medication alerts, we focused on high volume alerts with acceptance rate below 10%, defined as the number of medication orders that were cancelled, removed or discontinued within one hour after the alert fired. For BPAs, we focused on high volume alerts and reminders with cancellation rate above 95%. We performed a data-driven impact assessment and made changes through deliberation in the institutional CDS committee.

Results
The overall number of medication alerts shown was 3,563,354, with 27% seen by clinicians, with an average of 483 alerts per clinician per year. Number of alerts in groups determined by the volume and acceptance level is shown at the top of each bar (Figure 1). For example, 11 alerts with 5-10% acceptance accounted for 13% of alert volume. By switching off inactive allergy ingredient alerts, we reduced their volume by 24%.

The overall number of BPAs was 5,833,045, of which 36% were seen by clinicians, with an average of 980 BPAs shown per person per year. As an alert fatigue mitigation strategy we decided to turn off BPAs for "uninsured mammograms" and “Hepatitis C screening”, which entailed 2 of our top 5 BPAs by volume in the system.

Conclusion
Assessing the level of acceptance and appropriateness of alerts and reminders is an important task but it is often not performed due to insufficient resources and prioritization. We show here how a pragmatic and targeted approach focused on the highest volume alerts and reminders can be implemented with relatively limited resource allocation. Reducing alert fatigue remains a critical area of need for effectively supporting clinical decision making and improving patient care.

References
Towards the creation of a novel career-based Health Informatics (HI) curriculum assessment instrument: Mapping HI job competencies to HI curriculum competencies

Anand Kulanthaivel, MIS\textsuperscript{1}, Enming Zhang, MS\textsuperscript{1}, Shilpa Katta\textsuperscript{1}, Josette F. Jones, RN, PhD\textsuperscript{1}

\textsuperscript{1}Indiana University, Indianapolis, IN, USA

Introduction

In order to best determine what competencies and skills are required for various careers in Health Informatics (HI) and create appropriately matched academic curricula accreditation recommendations, it is important to inventory current HI industry job requirements and posted curricula outcomes with respect to existing curriculum assessment frameworks. For this study, a Clinical Informatics-related career competency list as published by the American Nursing Association (ANA)\textsuperscript{1} is used as a guiding framework to establish the competencies required in HI-related careers.

The Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM), in conjunction with the American Medical Informatics Association (AMIA), furthermore, have created a list of seven knowledge domains and skills relevant to HI curricula; three of these domains (behavioral/social sciences, health sciences, and information technology) form the core while the other four domains are conceptual overlaps of the three basic competency domains.\textsuperscript{2} Anderson & Krathwohl\textsuperscript{3} performed a review and revision of Bloom's classical taxonomy\textsuperscript{4} of learning; it is thus possible to utilize the taxonomy's revision in tandem with the CAHIIM-AMIA knowledge domain classification in order to assess the ANA’s instrument with respect to how its competencies match the competencies set forth in general pedagogy as well as those seen in existing HI curricula.

Methodology & Initial Results

Prior to coding, several changes to the provided terminologies were required as per consensus between the three producing authors (AK, EZ, SK). As needed, the titles and descriptions of the CAHIIM-AMIA competency domains were modified and definitions expounded upon in order to make them less ambiguous and more inclusive. Many competencies in the ANA instrument were re-worded by the authors in for similar reasons.

Two co-authors (EZ, SK) have determined which CAHIIM-AMIA competency maps best to each ANA job competency; furthermore completed is making the CAHIIM-AMIA domains more specific by mapping to the two dimensions of Anderson's revision of Bloom's taxonomy. Initial analysis of interrater reliability of this mapping shows that there is a raw concordance of over 60% between the two raters.

Conclusion & Future Work

We conclude at this point the fact that current assessment instruments require more explicit detail and other modifications in order to be useful. For validation reasons, a statistically sound method will be used to compare inter-rater reliability beyond raw percentage of concordance. Broader future goals include the creation of novel curriculum assessments that are based upon appropriate intelligence from both the academic world as well as that of the professional one.

References

2. The Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM). Competencies for Master's Programs in Health Informatics. (Webinar).
Participants’ View on Picture Taking in a Behavior Change Intervention

Pei-Yi Kuo, MA , Michael S. Horn, PhD
Northwestern University, Evanston, IL

Introduction

In diary studies and behavior intervention research, asking participants to take photographs of their actions is a common approach to elicit reflection1. Pictures can hold participants accountable2 and serve as source of learning and motivation3. Yet, we do not fully understand how users feel about responding to daily prompts by recording both words and pictures1. This paper presents qualitative findings from a 20-day behavior intervention on sustainable lifestyle choices and examines how people feel about capturing themselves using words and pictures.

Method

A total of 62 adult participants (83.6% female) were recruited for a 20-day intervention. The study involved asking participants to complete one sustainable lifestyle challenge per day (e.g., challenges related to exercise, diet, and transportation). Challenges were delivered to participants through a smartphone app called PACO4. Participants were required to describe what they had done for each challenge and submit an accompanying photograph. At the end of the intervention, participants filled out a post-study survey. We then selectively interviewed 15 participants who had disagreed with the survey questions: “I liked that I could capture my behavior with a picture.” and/or “The option to capture my behavior with the camera made me more aware of my behavior.” These participants were asked to reflect on how they felt about capturing their behaviors with words and photographs.

Findings

In the survey, the majority of participants indicated that they liked capturing their behavior with a picture (48.4%) and that this made them more aware of their behavior (46.8%). 67.7% of the total participants indicated they liked describing their behavior in the app. We found the male participants in our study had a positive attitude toward capturing their behavior with a picture, unlike earlier findings on the relationship between gender and photo sharing1. Similarly unexpected, only two out of 13 participants above 35 years old disliked the idea of capturing their behavior with a picture. On the one hand, when participants indicated that they disliked photographing their behaviors, their reasons included difficulty remembering to take pictures and difficulty capturing actions with photos (e.g., do 20 push-ups). Few participants said that taking pictures made the challenges more difficult and few did not like to be held accountable with pictures. On the other hand, some preferred describing what they had done with words only, saying that it seemed more helpful and logical, and that writing helped them think more.

Conclusion

To better understand the value of participant photographs in behavior change intervention, we examine what participants liked and disliked about capturing their behaviors with pictures. Though participants who liked capturing their behavior with a picture did not necessarily complete more challenges or describe their behaviors with more words, they mentioned that it was fun and they felt motivated. By talking to those who disliked capturing themselves, we identify several areas of improvement to make photographing behaviors a smoother process. These included: allow participants for more time to record pictures than words, enable separate recordings of words and pictures, and offer examples on capturing actions upfront. Our next steps will also involve developing a framework to determine the amount of user efforts spent in challenge completions from participants’ photos.

References

4. The personal analytics companions [Internet]. Available from: http://pacoapp.com
Integrating Usability Engineering into System Testing and Implementation: A Multi-Phased Approach to Ensure System Usability and Safety

Andre W. Kushniruk, MSc, PhD, FACMI1, Elizabeth M. Borycki, RN, PhD1
1University of Victoria, Victoria, British Columbia, Canada

Abstract
Usability has become recognized as being a major issue in the successful design and implementation of health information systems. We describe a multi-phased approach to integrating usability engineering into system testing to ensure both usability and safety of healthcare IT upon widespread deployment. The approach integrates laboratory testing with clinical simulations (conducted in-situ prior to widespread system deployment) and near-live recordings of user interactions with systems. At key stages in this sequence usability problems are rectified.

Introduction
Ensuring the usability and safety of healthcare IT is essential in order to lead to successful adoption of healthcare IT. Many different approaches have been proposed and applied for testing the usability of systems and for evaluating their impact on patient safety 1. While each method has its advantages, there are also limitations associated with each. For example, usability inspection methods can be easily and cost-effectively applied to make predictions about problems users may face. However, these methods are not a replacement for live testing of users involving application of usability testing methods and clinical simulations conducted in-situ 1. In this poster we argue for a multi-phased and mixed methods approach to ensuring system usability and safety. It is argued that in order to detect majority of errors, the methods should be applied in a phased approach, beginning with laboratory testing, followed by clinical simulations and near-live testing of systems.

Methods
An approach to the testing of healthcare IT was developed that involves several sequential phases. In the first phase classic usability inspection methods (including cognitive walkthrough and heuristic evaluation) are applied to predict usability problems. This is followed by laboratory-style usability testing under artificially controlled conditions, where video recording is used to collect data on typical user interactions. Problems arising from these phases are rectified and then a phase of testing under simulated conditions (i.e. by applying clinical simulation methods) is conducted. After rectifying any serious problems, the testing moves to the phase of near-live testing, where user interactions are observed under close to or real conditions and settings.

Results
The multi-method, phased approach described above has now been applied in a number of studies of EHRs and DSSs. The type of errors and usability problems detected at each phase were found to differ. For example, early phase usability inspection and usability testing typically uncovered surface level usability problems, but testing involving clinical simulations was needed to detect issues related to impact on clinical workflow. Near-live testing typically revealed a range of contextual issues that would otherwise have not been predicted by the previous phases.

Conclusion and Discussion
In our work we have attempted to integrate a number of different methods for ensuring system usability and safety. It was found that each method detected different types of problems, starting with surface level usability problems and leading to issues related to workflow and patient safety. It is argues that such a multi-method phased approach is needed to create a “safety net” to detect and rectify a range of errors and problems prior to widespread system release. The approach described in this poster has begun to be applied more widely in organizations where usability and safety have been deemed of utmost importance in system implementations and deployments.

References
Development and Validation of an Electronic Health Record (EHR)-Based Risk Stratification Rule for Inpatient Delirium

Joanne LaFleur, PharmD, MSPH,1,2 Jacob Crook, MStat,2,1 Scott Nelson, PharmD, MS,2 Lacey J Lewis,2,1 Kristin Knippenberg, MS,2,1 Grace Gardner, RN,2 Charlene R. Weir, RN, PhD, FACMI2,1

1University of Utah, Salt Lake City, UT; 2Department of Veterans Affairs, Salt Lake City, UT;

Introduction
Delirium during a hospital stay can have profound implications for elderly inpatients, increasing the risk of death at discharge and 12 months by 20% and 100%, respectively.1,3 Though reported rates vary from 14-56%,4,5 it is thought the true incidence is unknown because of low awareness and poor documentation.6 Though delirium-prevention strategies are effective,7 inpatient delirium remains largely unaddressed. We developed and validated a risk algorithm for identifying patients at high risk of delirium, with the goal of increasing clinician awareness of delirium risk in appropriate patients.

Methods
Our algorithm used structured data collected as a routine part of healthcare operations for a national cohort of hospitalized Veterans ages 40+ in 2011-2013. We developed the model using Cox Proportional Hazards Regression to predict delirium associated with an order for sitters or restraints. Candidate risk factors were clinical and demographic traits known or theoretically associated with inpatient delirium. We randomly sampled 8 x 50,000 patients from the full cohort and fit a model in each sample using backward selection. Risk factors that were significant at alpha=0.05 in at least 7/8 models were fit in the full cohort to estimate coefficients, which were multiplied by 10 to create “scores” for each risk factor. Scores generated for each patient based on the presence of risk factors. Test characteristics were calculated in the full cohort and in an external sample of prospectively assessed Veterans from one hospital. In the external sample, delirium was assessed using the Short Confusion Assessment Method (Short CAM).

Results
In 700,180 Veterans, 15,656 (2.2%) had orders for sitters or restraints. In the national cohort, accuracy for predicting sitter/restraint orders was maximized at a score threshold of 33 (C=73.1%). Test characteristics for the tool at that threshold are summarized in the Table. These results suggest that every 2.5 patients scoring as “positive” would include one case of true delirium. Higher thresholds (i.e., scores ≥175) are needed to obtain similar positive predictive values for predicting sitter/restraint orders.

Conclusion
Patients at highest risk for delirium can be identified using data collected as a routine part of healthcare operations. Future work should focus on developing and evaluating a clinical alert for delirium.

References
Integration of a Healthcare Data Dictionary Front-end Tool with Data Visualization: Architectural Features and Early Implementation in the Laboratory Domain

Jaehoon Lee, PhD1, Nathan C. Hulse, PhD2, Michael G. Newman, MS2, Douglas Mitchell1
Gyan Sharma, MS1, Naveen Maram, MD, MPH, MSHI1

1Intermountain Healthcare, Salt Lake City, UT; 2University of Utah, Salt Lake City, UT

Abstract

In this manuscript we integrated interactive data visualization with a web-based healthcare data dictionary (HDD) front-end tool. We used a commercial visualization tool: Tableau 1) to pipeline data extraction, aggregation, and visualization generation and 2) to use Tableau server as a visualization service provider. We implemented data visualization using the laboratory domain and the early usage in production demonstrated that the proposed approach is plausible in terms of computational efficiency and productivity of visualization pipelining.

Introduction

Athena is a locally developed web-based tool at Intermountain Healthcare used to support knowledge engineers to navigate HDD concepts and relationships. What the tool has not yet provided is information of clinical instance data to inform how HDD concepts have been practically utilized in Intermountain’s clinical data repository (CDR). The goal of the Athena visualization project is to enhance the tool with visual representations and interactive navigation of clinical data in conjunction with HDD concepts and terminologies.

Method

We used Tableau as a toolkit for whole data visualization pipelining, which includes 1) data retrieval and abstraction, 2) visualization transformation, and 3) visualization service. The visualizations were created in an authoring environment (Tableau Desktop) using data and knowledge extracted from HDD and CDR. Once the visualizations were created and deployed upon a Tableau server, the server provided remote retrieval as a service based on Tableau Javascript APIs, such that Athena could embed the visualizations in its web pages. Access control and user privileges in the two applications were synchronized so that patient health information could be managed in a secured way. The visualizations on the Tableau server side constantly refreshes their data to maintain an up-to-date status in an automated manner separately from Athena.

Conclusion

The implementation in production was successful and its early usage pattern demonstrated that 1) the integration of the two applications are feasible in terms of architecture such as interfaces, security control, and performance; 2) data pipelining was separated from the Athena application and fully managed on Tableau side to ensure productivity and maintainability of the visualizations, and 3) visualization response time was acceptable (average elapsed time to view: 0.55 seconds / request). Our future plan is to add diverse use cases and domains based on the framework such as problems and medications.
Clinical Decision Support for High Risk Medications in the Elderly at the Point of Prescribing

Joseph R. LeGrand, PharmD1, Erin B. Neal, PharmD2, Bryan E. Shepherd, PhD3, Scott D. Nelson, PharmD, MS1,3, Shane P. Stenner, MD, MS1,3

1Vanderbilt University Medical Center, Nashville, TN, 2Vanderbilt Health Affiliated Network, Nashville, TN, 3Vanderbilt University, Nashville, TN

Introduction
High risk medications (HRMs), such as those with high anticholinergic and antihistaminergic activity are associated with increased morbidity and mortality in the elderly.1 The American Geriatrics Society (AGS) recommends avoiding HRMs in older adults to improve safety and quality of care in this population.2 Further, performance on HRM metrics now impact reimbursement of health systems, payers, and pharmacies as national quality organizations have adopted these recommendations. Several interventions have targeted HRM prescribing in the elderly with inconsistent results, and only a few have incorporated clinical decision support (CDS) at the point of prescribing.3,4

Methods
The most common HRMs prescribed using Vanderbilt’s e-prescribing system were evaluated in order to target the CDS rules for the most frequently used medications. Targeted CDS was limited to those HRMs that had at least one evidence-based potentially safer medication (PSM).5 A PSM alert is triggered within the e-prescribing workflow when a HRM prescription is initiated for a patient 65 years or older. The alert displays a list of PSMs, as well as an option for the user to continue with the current HRM. If a PSM is selected, the user is provided available doses and dosage forms and can continue with the order. Descriptive statistics were collected for the target population, and post-implementation rates were compared to baseline rates using a quasi-Poisson regression model in R to detect a change in prescribing habits.

Results
Following the implementation of a HRM CDS alert, a 10.6% decrease (95% CI 6.5% to 14.5%, p < 0.0001) in the risk of prescribing HRMs in the target population was observed.

Table 1. Average weekly percent and quantity HRM in target population pre- and post-intervention.

<table>
<thead>
<tr>
<th>12 Week End Date</th>
<th>6/27/2015 (pre)</th>
<th>9/19/2015 (pre)</th>
<th>12/12/2015 (pre)</th>
<th>3/5/2016 (post)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg % HRM / Week</td>
<td>3.01%</td>
<td>3.01%</td>
<td>2.96%</td>
<td>2.70%</td>
</tr>
<tr>
<td>Avg # HRM / Week</td>
<td>265.3</td>
<td>255.6</td>
<td>265.1</td>
<td>235.5</td>
</tr>
</tbody>
</table>

Conclusion
CDS at the point of prescribing is effective at reducing the incidence of HRM prescribing in the elderly. It is unknown if the reduced prescribing of HRMs results in decreased adverse drug events or improved patient outcomes. One limitation of the study is that a control group was not established, which could have helped identify any potential confounding factors.

References
Designing Avatars to Assist with Self-Management for Overweight and Obese Adolescents

Cynthia LeRouge, PhD¹, Savitha Sangameswaran, MS¹, Toree Malasanos M.D.²
¹University of Washington, Seattle, WA, ²Kiwee Health, LLC, Gainesville, FL

Introduction
Adolescent overweight and obesity is a major public health concern. Evidence has accumulated about the efficacy and user acceptance of computer-based interventions (CBI) to modify behaviors. Given the interest that youths have in avatars from their experience with avatars in computer games, consumer health applications designed with virtual agents (a.k.a. avatar characters representing the user in virtual environments or artificial health assistant characters) may be an effective way to engage them in behavioral lifestyle support tools. The objective of this study is to examine adolescent preferences regarding avatar design and functionality (i.e., the collective features and functions adolescents will find most useful, usable, empathetic, and sociable) for potential use in a healthy lifestyle/weight loss assisting application.

Methods
Using one-on-one user-centered design methods, 77 adolescents (ages 12 to 17) enrolled an intensive lifestyle modification summer camp program served as study participants. The adolescent participants: 1) responded to general questions of computer use (i.e., technology access questions, level of avatar, or virtual agent experience) 2) reviewed mid-fidelity mock-ups of seven types of graphical embodiments of virtual agents and provided design recommendations regarding specific attributes of the character(s), such as voice (options presented from professional recordings), accessories available to the character, and a facial expressions chart review and 3) provided data on the interactive functions the avatars could serve. Thematic analysis of the transcripts was conducted by a team of three researchers.

Results
Adolescents indicated a great level of interest in avatars assisting self-management efforts – “…Yeah I think it would be really cool to like be able to make a creatable Avatar, and see like maybe have progress shown in the Avatar.” Some of the adolescent preferences for avatar design are outlined in the Table below.

<table>
<thead>
<tr>
<th>Avatar Design</th>
<th>Adolescent Preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice</td>
<td>Teen or adult voice that was empathetic and “softer”</td>
</tr>
<tr>
<td>Facial Expression</td>
<td>Range of emotions desired included cautious, disappointed, frustrated, confused, curious, grieving, confident, optimistic, sad, surprised, horrified, funny/silly, and sick.</td>
</tr>
<tr>
<td>Character types</td>
<td>Ranged from cartoon to realistic, human based avatars including representation of self</td>
</tr>
<tr>
<td>Avatar Accessories</td>
<td>Sunglasses, headphones or ear buds, gadgets, necklaces, musical instruments, water bottles, football/sports related memorabilia and objects, and self-created tattoos</td>
</tr>
<tr>
<td>Reward System</td>
<td>Obtain new/ additional accessories, clothing when goals achieved</td>
</tr>
<tr>
<td>Avatar Contextual Environment</td>
<td>Range included jungle, forest, a classroom setting, grocery store, small town environment, track/field environment, football field, basketball court, chef's kitchen, and a wizard's room</td>
</tr>
<tr>
<td>Number of Avatars</td>
<td>Minimum of one avatar. Strong preference for two - one as coach other would represent the user.</td>
</tr>
<tr>
<td>Gaming with Avatar</td>
<td>Single Player and Multi-Player Games include Food Choices, Portions Sizes, Cooking Simulation</td>
</tr>
<tr>
<td>Display</td>
<td>“Virtual Me” avatar exemplifies goal progress through both external avatar appearance &amp; internal changes of the body (e.g., a beating heart that reflect the effects of progress &amp; choices)</td>
</tr>
<tr>
<td>Interaction</td>
<td>Ranged from initial greeter to exercise pal to always present empathic support</td>
</tr>
</tbody>
</table>

Conclusion
Our results call for a mass customization (with options such as hair and eye color) approach to the design of avatars that can facilitate self-management of overweight/obese adolescents and augment efforts in the continuum of care. Designing avatars based on adolescent preferences may increase the probability of engagement and long-term retention of participants with behavioral lifestyle support tools.
Reviewing Asthma-related Grey Literature and Personal Opinions on Twitter using LDA and CTM Clustering

Gondy Leroy, PhD, Joe Koolippurackal, Shikha Swami, and Philip Harber, MD MPH
University of Arizona, Tucson, Arizona, USA

Abstract
Twitter is a big data source useful for sampling opinions and information in the community. In this exploratory study, we compared the usefulness of information shared about asthma. We collected tweets using asthma and #asthma and compared tweets with and without URLs. To automatically create topic overviews that can be efficiently reviewed, we used LDA and CTM clustering. We evaluated using 2, 5 or 10 clusters.

Introduction
Social media play an increasingly important role in medicine. We focused on asthma, which affects 8% of the US adult population(1), and analyzed Tweets to assess: (a) what proportion is unique; (b) differences between tweets from individuals vs grey literature (i.e., those referring to professional information); (c) whether searching for tweets with or without a hashtag matters.

Methods and Results
In fall 2016, we collected 178,796 tweets for asthma and #asthma using the Twitter Search API. Based on prior work(2), we made the broad assumption that tweets with a URL refer mostly to grey literature tweets (by professionals), while tweets without URL are more often comments by individuals. We removed duplicate tweets and stopwords, then stemmed the terms. To review tweet content, we compared Latent Dirichlet Allocation (LDA) and Correlated Topics Model (CTM) clustering for 2, 5 and 10 cluster.

Overall, 70% of the tweets were unique (Table 1). Tweets with a URL are 64-77% unique. Tweets without URLs were more unique and almost completely unique for asthma. This is not unexpected since this query is broad (appearance of word asthma) and personal comments can be expected to be unique. To evaluate the content of the entire set in an efficient manner, we clustered the tweets and evaluated the clusters. All cluster sets were shown in random order and evaluated (blind) by a domain expert, who scored each term on a 4-point scale of relevance to asthma (an asymmetric scale was used to penalize nonsense terms). Figure 1 and 2 show the average normalized scores. Scores for all terms were combined per condition. LDA clustering resulted in more relevant clusters and more consistent quality. Tweets for #asthma (with hashtag) by individuals contained the most relevant terms. Terms were least relevant for the asthma query without URL (personal comments, broad query). Implication: For estimating community members’ opinions & information, relevance is improved by use of LDA clustering of tweets without URL references.

Table 1: Tweets Collected

<table>
<thead>
<tr>
<th></th>
<th>all / unique (% unique)</th>
<th>asthma</th>
<th>#asthma</th>
</tr>
</thead>
<tbody>
<tr>
<td>With URL</td>
<td>27,054 / 17,227 (64)</td>
<td>52,508 / 40,219 (77)</td>
<td></td>
</tr>
<tr>
<td>Without URL</td>
<td>5,579 / 4,252 (76)</td>
<td>66,601 / 64,320 (97)</td>
<td></td>
</tr>
</tbody>
</table>

References
Drug-drug Interaction Detection with A Topic-modeling Based Framework Augmented with Distant-supervision

*Dingcheng Li1, *Sijia Liu2, Majid Rastegar-Mojarad1, Yanshan Wang1, Xiaodi Li3
Vipin Chaudhary2, Terry Therneau4, Hongfang Liu1
1 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
2 Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
3 Department of Mechanical Engineering, Donghua University, Shanghai, China

Introduction Potential drug-drug interactions, defined as the co-prescription of two or more drugs that are known to interact, are one of primary causes of medical error. A report in Australia mentioned that 75% of hospital admissions related to medication errors are preventable. Indeed, health care providers often have inadequate knowledge of what drug interactions can occur of how to properly manage an interaction when patient exposure cannot be avoided. One of the primary reasons for this knowledge gap is the lack of complete and authoritative source of DDI knowledge. The dynamic nature of drug knowledge, combined with the enormity of the biomedical literature, makes the task of collecting and maintaining up-to-date information on drug-drug interactions extremely challenging and time-consuming. Therefore, there is a strong need to approach this task with automated methods, supplemented with human effort. Natural language processing (NLP) and information extraction methods for identifying and extracting DDIs have been receiving increased attention in the last few years, and several attempts have already been made to develop methods for this task, showing good potential for success.

Methods In light of this, we propose to design, implement and evaluate a Bayesian model complemented with knowledge-driven distant supervision called relation topic modeling (RelTM for short). It is a semi-supervised probabilistic generative model for inducing clusters of DDI types and recognizing drug entities as well. The set of relation types is not pre-specified but induced from observed unlabeled data. The relation induction is augmented with distant supervision by incorporating resources-based similarity functions. Our observed data consists of a corpus of sentences and each sentence is represented by a bag of relation pairs. Each pair represents an observed drug-drug interaction between two drugs and consists of three components: the relation features, the source drug and the destination drug. The relation pair is the primary observed random variable in our model and we construct our models so that clusters consist of textual expressions representing the same underlying relation type. Since unlike classical LDA where observed words are independent and shared by the whole collection of documents, it will be reasonable to assume that each word may be generated by some specific topic. In our framework, a set of drug mention features supports one drug and relation features and mention features jointly support one relation. Therefore, the generative process is somewhat different. Namely, all of its features are only sampled under one specific relation and all of mention features are only sampled under one specific drug entity as well.

Table 1 Performance evaluation in DDI Drugbank dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Number of pairs</th>
<th>RelTM baseline</th>
<th>RelTM+SMC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>Mechanism</td>
<td>827</td>
<td>0.325</td>
<td>0.302</td>
</tr>
<tr>
<td>Effect</td>
<td>1700</td>
<td>0.495</td>
<td>0.536</td>
</tr>
<tr>
<td>Advise</td>
<td>1322</td>
<td>0.386</td>
<td>0.210</td>
</tr>
<tr>
<td>Int</td>
<td>188</td>
<td>0.135</td>
<td>0.526</td>
</tr>
<tr>
<td>Micro-average</td>
<td></td>
<td>0.335</td>
<td>0.393</td>
</tr>
</tbody>
</table>

Experimental Results As a gold standard for all experiments in this work, we used DDIExtraction-2013 corpus provided by the challenge organizer, for development. It contains 142 Medline abstracts on the subject of drug-drug interactions, and 572 documents describing drug-drug interactions from the DrugBank database. The corpus includes 6976 sentences that were annotated with four types of pharmacological entities and four types of DDIs. To demonstrate the effectiveness of our proposed model, we conducted a series of experiments on DDI Drugbank dataset. There are two models evaluated in this section: RelTM model and the combined model with RelTM and similarity measure coefficient (SMC) model. The performance metrics is computed by the code in Java and it is provided by DDIExtraction 2013 Challenge organizers. The performance of the two models mentioned above is shown in Table 1. Comparing these results, we can draw the conclusion that the combination of RelTM and distant supervision together improve the overall performance, which is indicated in the micro-average F1 score, from 0.362 increasing to 0.480.

Conclusion and Future Work In this work, we propose a topic modeling based distant supervised approach for the task of drug-drug interaction from biomedical text. The proposed approach does not require human efforts such as annotation and labeling in data preparation stage, which is its advantage in trending big data applications. Meanwhile, the distant supervision combination allows us to incorporate rich existing knowledge resources provided by domain experts. In future, we will extend similarity function by incorporating updating rules and correlation between latent relations. One direction of exploration includes the release of symmetric Dirichlet prior to non-symmetric considering the unbalanced numbers of relation categories.

Acknowledgements The authors gratefully acknowledge the support from the National Institute of Health (NIH) grant 1R01LM011934 and 1K99LM012021-01A1.

* These authors contributed equally to this study.
How Big Is Big? Amount of Data in Intensive Care Units DataMart Platform
M. Li, M.D., I. Tiong M.S., J. Dyke, L. Fan M.S., S. Plamadeala, V. Fedosov, M.D., Ph.D.,
B. Pickering, M.B.B.Ch., V. Herasevich, M.D, Ph.D.
Mayo Clinic, Rochester, MN

Introduction. Along with the widespread use of advanced information technology in the healthcare system the availability of a huge and heterogeneous amount of electronic medical data has proportionally increased. It has provided a novel opportunity to accelerate progress on the quality of patient care (1). Meanwhile, it also makes patient care more complex (2). There is no common clarity on how much data is being generated by electronic medical records (EMR), particularly in a highly data intensive setting environment like an intensive care unit (ICU) and how to manage the huge data amount for quality and research purpose. Mayo Clinic Rochester ICUs have an average of 15,000 admissions each year and generate considerably large amounts of clinical and administrative data stored and distributed in heterogeneous EMR systems.

The ICU DataMart at the Mayo Clinic is a near-real time database storing and integrating ICU patients’ clinical and administrative data which has served as a central data warehouse for decision support, quality control reporting, and data mining.

Objective. The aim of the current project is to quantify and understand the amount of data generated by ICU patients.

Methods. The Mayo ICU DataMart is a MS SQL based relational database integrating ICU patient clinical and administrative data extracted from various EMR systems and populated through extract transfer load (ETL) processes to DataMart in nearly-real time. We calculated the amount of one year of data which were generated during the year 2014.

Results. A total number of 13,355 patients with 19,409 ICU admissions were stored in Mayo Clinic Rochester ICUs in 2014 (summarized below).

<table>
<thead>
<tr>
<th>Data Points</th>
<th>Data Description</th>
<th>Total Records</th>
<th>Average Records/ Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vital Signs</td>
<td>Entire data points &gt;150 variables under monitor for all the ICUs</td>
<td>247,254,039</td>
<td>6,544</td>
</tr>
<tr>
<td>Fluid-in/Fluid-out</td>
<td>IV in, urine out, oral-in/out, tube-out</td>
<td>5,270,604</td>
<td>151</td>
</tr>
<tr>
<td>Laboratory Test</td>
<td>Entire lab test copied over from EMR System – SOFT for all the ICUs</td>
<td>3,640,762</td>
<td>277</td>
</tr>
<tr>
<td>Administered Medication</td>
<td>All medication records taken by IV, oral and shots for all the ICUs</td>
<td>546,682</td>
<td>42</td>
</tr>
<tr>
<td>Medication Order</td>
<td>All medication records copied by Enterprise Order system</td>
<td>1,114,392</td>
<td>85</td>
</tr>
<tr>
<td>Nurse Flow Sheet</td>
<td>Flow sheet chart for all the ICUs by nurse</td>
<td>22,899,994</td>
<td>1,697</td>
</tr>
<tr>
<td>Clinical Notes</td>
<td>Admission, discharge, problem list, impression, family history</td>
<td>6,759,504</td>
<td>521</td>
</tr>
<tr>
<td>Microbiology</td>
<td>Culture results and susceptibilities</td>
<td>137,959</td>
<td>23</td>
</tr>
</tbody>
</table>

Conclusion: Even though it is challenging to store a huge quantity of data and make it accessible, we have found that the amount of data generated by an ICU practice in an academic medical center is manageable and easily accessible with a customized ICU DataMart platform.

References.
Ontology-Based Analysis of Adverse Drug Reactions Associated with Anti-infection Drugs in China

Mei Li, MD1, Yuying Cao, MD1, Liwei Wang, MD, PhD1,2, Hongfang Liu, PhD2, Yongqun He, PhD3

1School of Public Health, Jilin University, Changchun, Jilin, China; 2Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; 3 Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract

To understand the characteristics of adverse drug reactions (ADRs) associated with anti-infection drugs in China, based on the Ontology of Adverse Events (OAE), we analyzed 2,355 ADR case reports from the literature including the patients’ age and gender distribution, the administration routes, the drug classes, the most and severe ADRs, and the characteristics of ADRs caused by drugs of different pharmacological effects.

Introduction

In China, the number of adverse drug reactions (ADRs) reports associated with anti-infection drugs (AIDs) has continuously ranked top according to the Annual Report on Adverse Drug Events Surveillance (2014)1. In order to avoid severe ADRs, it is important to understand characteristics of ADRs associated with AIDs. However, there is no existing way to systematically analyze the ADRs due to the lack of standards in Chinese for ADR terms. Recent research shows that the Ontology of Adverse Events (OAE)2, with complete logical reasoning and classification structure, describes concepts and the concept relationship in the field of ADRs. And it gives a clue for our study. This study aims to analyze the ADR reports from the literatures based on the OAE, in terms of the patients’ age and gender distribution, the administration routes, the drug classes, the top and severe ADRs, and the characteristics of ADRs caused by drugs with different pharmacological effects.

Methods

The method contains four steps: data collection, data cleaning, data standardization and statistical analysis. In the first step, we collected ADR reports associated with AIDs from “National Scientific Data Sharing Platform for Population and Health” (from the literature in China during 1996-2014). In the cleaning up step, we removed the data irrelevant to anti-infections drugs, and extracted ADR terms from reports. During data standardization, the drugs were classified according to the Chinese textbook New Pharmacology (17th version), and then ADR terms were mapped to OAE terms from Chinese to English. In the last step, we used the Excel and CIMMiner to retrospectively analyze ADRs in the case reports.

Initial Results and Discussion

In all ADR reports, ADRs of AIDs in male and female were basically similar. Most reports of ADRs were found among the patients aged 30 to 40 (18.76%). The most reports involving Traditional Chinese Drugs were for children under 10 years of age (23.81%). Intravenous drip is the administration route that caused the most ADRs. The drug category that caused the most ADRs was quinolone class. And the chest discomfort was the primary ADR in all reports. In total 31 drugs, corresponding to 9 drug classes, and 2 Traditional Chinese Medicines caused serious ADRs. The drug Fosfomycin Sodium for Injection has caused severe ADR of death.

In this study we used OAE to systematically and effectively analyze the characteristics of ADRs in Chinese, and discover the correlation between the drug category and ADR category. Building a bilingual Chinese-English OAE ontology is significant for analyzing ADRs based on existing ontology.

References

An Exploratory Study of Continued Use of HIE

Huigang Liang, PhD1, Yajiong Xue, PhD1, Anne Marie Robertson2, Yvonne Hughes, MPA2, P. Allen Gray, PhD2, Patty Lewis2

1East Carolina University, Greenville, NC, USA; 2CoastalConnect Health Information Exchange, Wilmington, NC, USA

Introduction

Health Information Exchange (HIE) can potentially improve the quality, efficiency, and safety of health care delivery and be useful in engaging patients and improving care coordination and public health surveillance. Although the number of HIEs in the U.S. is growing rapidly in recent years, many struggle to sustain a healthy business. According to a report from eHealth Initiative, many new HIEs are starting up and existing ones are seeking ways to overcome various hurdles. The use of HIE is low. A review study based on 85 HIE evaluation articles finds that most studies reported HIE use in only 2-10% of the patient encounters (Rudin et al. 2014). The actual impact of HIEs on healthcare industry is still unclear. Only 57% of published analyses reported HIE benefits (Rahurkar et al. 2015). The number of HIEs that have been evaluated is low. Only 7-10 out of more than 100 HIEs were evaluated (Rudin et al. 2014). Given the huge federal investments on HIT and HIE, it is important to understand what lead to physicians’ positive perceptions and continued use of HIE. The logic is that if physicians perceive that HIE is valuable and are satisfied with HIE, they will be more likely to continue using HIE, thus reducing the sustainability problem of HIE.

In this research, we develop a theoretical model based on the IS success model (DeLone & McLean, 1992; 2003) to explain what factors influence continued use of HIE. We propose that the benefits of HIE are mainly derived from three key characteristics: data quality (accuracy, reliability, completeness, and timeliness of HIE data), data accessibility (the degree to which external health data can be accessed via HIE), and system compatibility (the degree to which HIE is compatible with the users’ current system and work context). These three factors can improve operational efficiency by reducing work time, costs, and test redundancy. Improved operational efficiency and data quality will help to create a heightened value of HIE. Further, if HIE is perceived to increase efficiency and bring value into the practice, a greater level of user satisfaction will be achieved. Finally, we propose that perceived value, operational efficiency, and user satisfaction will jointly enhance the continued use of HIE.

Method

Coastal Connect Health Information Exchange (CCHIE) is a nonprofit regional health information organization dedicated to enabling the secure and reliable exchange of health information in Eastern North Carolina. Supported by a Duke Endowment Grant, CCHIE engaged in a three-year project to connect Community Care of North Carolina (CCNC) providers, safety-net organizations, and some specialists identified by CCNC through a HIE network. By electronically sharing patient data, CCHIE intended to increase the efficiency of patient care and improve care coordination. This study is part of a large project to evaluate the outcome of the HIE implementations in participating physician practices from 2012 to 2014. For this study, we surveyed 626 managers of physician practices that joined the CCHIE network. A total of 126 completed surveys were returned. We used linear regression to estimate the relationships among variables.

Results

As hypothesized, operational efficiency is increased by data quality (b=.33, p<.01), data accessibility (b=.31, p<.01), and compatibility (b=.23, p<.01). Perceived value is influenced by data quality (b=.24, p<.05) and operational efficiency (b=.48, p<.01). User satisfaction is enhanced by perceived value (b=.45, p<.01) and operational efficiency (b=.42, p<.01). Finally, Continued use is increased by satisfaction (b=.28, p<.05) and operational efficiency (b=.18, p<.05), but not perceived value (b=.15, p>.05). Further analysis shows that the effect of perceived value on continued use is fully mediated by satisfaction. The explained variance (R²) for efficiency, value, satisfaction, and continued use are .45, .63, .62, and .31, respectively. Common method bias was addressed by the Harmon’s one factor test.

Conclusion

Continued HIE use can be directly promoted by user satisfaction and operational efficiency, and indirectly promoted by perceived value, data quality, data accessibility, and system compatibility.

Reference (Available upon request)
Three-Dimension (3D) Blood Vessel Reconstruction and Spatial Analytics with Whole-Slide Histological Images

Yanhui Liang, MS1, Jun Kong, PhD2, Fusheng Wang, PhD1
1Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY; 2Department of Biomedical Informatics, Emory University, Atlanta, GA

Abstract

Digital pathology images present to researchers and clinicians rich information on tissue morphological and pathological signatures at the cellular level. In this study, we report a complete framework for 3D histological structure reconstruction of blood vessels with serial whole-slide microscopy images, and a high performance spatial query system to characterize the spatial relationships across reconstructed 3D histological objects. We vision our work as a new avenue to facilitate biomedical research with histopathological imaging data.

Introduction

Thanks to advances of large-throughput scanning technologies, digital pathology images are increasingly adopted in clinics for disease diagnosis and treatment evaluation. By studying whole-slide histological images, researchers and clinicians can better understand the underlying mechanisms of disease onsets and pathological evolutions. Therefore quantitative imaging and spatial analysis of histological entities with microscopy images are in great demand, especially for those studies involving 3D histological structure characterization and 3D spatial analytics (such as blood vessels and nuclei). We present such a framework for 3D primary vessel reconstruction with a set of serial histological images, and a high performance data warehousing system for large-scale spatial analytics with 3D medical data.

Methods

The Framework for 3D Blood Vessel Reconstruction (Figure 1): Our 3D vessel reconstruction framework consists of multiple processing components: image registration, vessel segmentation, vessel cross-section association, intermediate slide interpolation and 3D volumetric rendering. Given a set of histological whole-slide images, we register all slides to a reference image to align all slides into the same space. We next segment vessels by an improved level set method with prior information on vessel wall probability in the energy minimization paradigm. We achieve the optimal vessel cross-section associations by local bi-slide mapping and global vessel structure association within a Bayesian Maximum A Posteriori (MAP) framework. We generate intermediate slides by B-spline interpolation for smooth transition, and finally visualize the reconstructed 3D primary vessels by volumetric rendering with mesh model.

The High Performance System for 3D Spatial Queries (Figure 2): As changes of 3D histological structures and their spatial relationships are essential to study various diseases, we build a high performance 3D data warehousing system to store, manage and analyze 3D histological objects. The system supports multiple spatial query cases, such as 3D spatial cross-matching and 3D spatial proximity query. Spatial cross-matching of reconstructed 3D histological objects (vessels or nuclei) from different methods can be used to evaluate algorithm performance, and spatial proximity query such as nearest neighbor can inform pathologists of such answers as “for each 3D stem cell, return the nearest 3D blood vessel and its distance”. Our system includes 3D data partitioning, a real-time 3D spatial query engine and MapReduce based query processing. It is scalable and effective for querying massive 3D spatial data.

Conclusion

We evaluate our 3D reconstruction framework with a set of 54 whole-slide images from liver tissue sections stained by Immunohistochemistry (IHC), and apply our 3D spatial query system to the reconstructed 3D histological objects. Experiments present satisfactory results and quantitative evaluations demonstrate the efficacy of our methods. Our work is generic and applicable to other 3D biological entities with 3D pathology image datasets of diverse diseases.

References

Abstract

Team Based Learning (TBL) consists of several well established components: individual readiness assurance test (iRAT), team readiness assurance test (tRAT), team applications (tAPP), and peer evaluation. We have developed integrated web-based modules that deliver the TBL components to learners via the web. Facilitators are able to monitor on going and cumulative results.

Introduction

Team Based Learning (TBL), which is student centric & accountable and which is used in many medical schools, consists of several well established components: individual readiness assurance test (iRAT), team readiness assurance test (tRAT), team applications (tAPP), and peer evaluation. In order to streamline the delivery of TBL components to learners in a paperless fashion and to simplify facilitator administration of TBL these components were made into web deliverable modules.

Technology

A web services solution stack consisting of Apache web server, MySQL database and PHP (hereinafter, web stack) was used to develop the TBL delivery system. Initially Uniform Server and later Bitnami web stacks were employed. The Uniform Server, a Microsoft Windows operating system web stack, can run from a thumb drive or any USB storage device and requires no installation. Thus the system can simply be used on any computer with Windows operating system and a static IP address. The Bitnami web stack can run in Windows, Mac OS X or Linux environments as well as VMware or VirtualBox virtualized environments, and popular cloud platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform.

Design

The tRAT & tAPP components have the full functionality of expensive self-scoring scratch off multiple answer sheets. Real time monitoring show facilitators each team’s correct & wrong answers, duration time and completeness as well as team test scores for each tRAT & tAPP and cumulative scores and rankings. The iRAT module allow only one answer attempt and provide facilitators with the same metrics as the tRATs and tAPPs. The Peer review modules allow each team member to rate their peers using either radio buttons, sliding bars, or divide the points as well adding text comments. The peer evaluations are available anonymously to team members & non-anonymously to facilitators from their respective home pages. Midterm and final grades are automatically calculated using any percent weighting for iRAT, tRAT, tAPP, and peer evaluation.

Conclusion

This integrated web based system delivers all of the necessary TBL components in an efficient and paper & printer saving manner that is functionally consistent with the paper format. When asked anonymously their preference, 8 students preferred web based delivery, 5 students preferred paper based, and 3 had no preference. Modules monitor, in real time component results and progress. Database storage of learner data makes for rapid summative assessments.
A Comparison of ESpell and GSpell Spell-Checker Tools

Fang Liu, MS; Paul Fontelo, MD, MPH
National Library of Medicine, Bethesda, MD

Abstract
Accurate spelling is crucial for obtaining optimal results with PubMed search tools. With increased usage of mobile devices and smartphones, we have noticed higher spelling errors. Auto spell-checkers, like, ESpell and GSpell, provide spelling suggestions and improve data retrieval. Using actual 500 search terms, we compared the use of these two utilities in ‘PubMed for Handhelds’. Both utilities are useful but we found that ESpell has higher precision and recall.

Introduction
ESpell and GSpell spell-checkers, both developed at National Library of Medicine (NLM), provide spelling suggestions to improve retrieval of PubMed data. ESpell, a recent addition (2015), is part of NCBI’s E-Utilities suite and is implemented through an API in existing search tools to provide spelling suggestions for misspelled terms within a single text query. GSpell uses several word similarity algorithms and can provide multiple suggestions even with compound words. Its Java software requires software installation on a server and migration when servers are updated or replaced. GSpell development and support ceased in 2006, so it is made available under "as is" condition. GSpell was used in “PubMed for Handhelds” for many years, but a recent hardware upgrade made migration challenging. Our goal for this project is to compare the ESpell and GSpell auto-suggestion functions.

Methods
From askMEDLINE’s query database, 500 search queries which retrieved no citations were randomly selected. The set of queries were fed to a script which simulated the automatic spelling correction function for both ESpell and GSpell. GSpell was configured to return the closest suggestion only. A manual normalization process removed queries that were Non-English, numbers (PubMed IDs) or people names, since spell-checkers are unable to autocorrect these terms. One of the authors (PF) evaluated the spelling suggestions to determine the accuracy of suggestions.

Result
Of the 500 queries, 45 (9.0%) were non-English queries, 66 (13.2%) were person names, and 11 (2.2%) were numbers. There were 378 queries in the normalized set. Among the 378 queries remaining, 175 were misspelled terms, and 203 queries were correct terms. ESpell gave suggestions for 124 misspelled queries (109 correct) and on 1 correct query. GSpell had autosuggestions on 127 misspelled queries (71 correct) and on 1 correct input query (“Dimexidum”). The precision for ESpell and GSpell are 87.2% (109/125) and 55.5% (71/128), respectively and their recall rates are 62.3% (109/175) and 40.6% (71/175), respectively.

<table>
<thead>
<tr>
<th>Total</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-English terms</td>
<td>45 (9.0%)</td>
</tr>
<tr>
<td>Name</td>
<td>66 (13.2%)</td>
</tr>
<tr>
<td>PubMed ID or Number</td>
<td>11 (2.2%)</td>
</tr>
<tr>
<td>Normalized input</td>
<td>378 (75.6%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Misspelled Terms</th>
<th>ESpell</th>
<th>GSpell</th>
</tr>
</thead>
<tbody>
<tr>
<td>With autosuggestion</td>
<td>124</td>
<td>127</td>
</tr>
<tr>
<td>No autosuggestion</td>
<td>51</td>
<td>48</td>
</tr>
<tr>
<td>Correct suggestion</td>
<td>109</td>
<td>71</td>
</tr>
<tr>
<td>Input Correct (203)</td>
<td>ESpell</td>
<td>GSpell</td>
</tr>
<tr>
<td>Autosuggestion</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Left panel, normalized data; middle panel, autosuggestion comparison; right panel, precision and recall.

Conclusion
We found both spell-checkers useful but from these results, ESpell seems to show better precision and recall. A recent hardware upgrade also provided challenges in GSpell installation. ESpell is now integrated in all “PubMed for Handhelds” servers.
Implementing Precision Predictive Analytics for Unplanned 30-Day Readmissions for Real-Time Use at the Point of Care

David F. Lobach, MD, PhD, MS1; Deb McGowan, BSN2; Nish Hartman, BA3; Michael Dulin, MD, PhD4; Marcy Neale, MS4; Jacqueline Mikuleza, BSN4; and Jean Wright, MD, MBA4

1Klesis Healthcare, Durham, NC; 2COPD Foundation, Washington, DC; 3Predixion Software, Aliso Viejo, CA; and 4Carolinas HealthCare System, Charlotte, NC

Introduction
Risk prediction models are increasingly used to identify patients who are likely to experience high cost health events, especially hospital readmissions.1 While these models have variable predictive capacity, the results have had only limited impact on readmission rates.2 One issue that may limit the impact of these predictive models on care outcomes is the lack of tools to enable the predicted insights to effect care delivery.3 In this project we solved this “last mile” problem by creating tools that integrated the results from a predictive readmission risk model into the workflow in order to use these insights to directly impact the delivery of care around the time of discharge. Identification of patient-specific modifiable factors allows personalized customization of the readmission prevention program though interventions that address specific patient care needs.

Methods
The model to predict the risk of unplanned 30-day readmissions was created using data extracted from electronic health records and billing data at Carolinas HealthCare System in Charlotte, NC. The data fields included: social-demographic characteristics, comorbidities, index admission and history of health care utilization, vital signs and other physical parameters, lab data, medications, and treatments/procedures. The target outcome for the model was defined as 30-day unplanned all-cause readmission that was aligned with the methodology used by CMS. We extracted 155,907 hospital discharges spanning two years and over 200 variables. The original data were randomly divided into a 70% training set and a 30% testing set. These data were analyzed using logistic regression for evaluating individual continuous covariates and chi-squared tests for categorical covariates. The modeling was conducted using the PredixionInsight tool (Predixion Software, Aliso Viejo, California), which allowed for multiple modeling techniques to be assessed (Naïve Bayes, Neural Network, Decision Tree, and Logistic Regression) and compared to one another for predictive accuracy.

Results
The readmission risk prediction model with the best performance used neural networks and retained 71 variables. It generated a C-Statistic of 0.78, sensitivity of 64.5%, specificity of 78.9%, positive predictive value (PPV) of 27.0% and negative predictive value (NPV) of 94.6%. Patient-specific predicted risk results were generated for all admitted patients and were refreshed hourly throughout the duration of a patient’s hospitalization. This information was communiccated to the care managers, physicians and other providers through a window inserted into the EHR that presented the readmission risk score within an individual patient’s record and through a dashboard that displayed risk scores for all patients on a hospital floor. This display included: 1) a list of patients with room number tagged with a colored flag indicating level of risk (low, medium, high, very high); 2) a percentage of patients within each category; and 3) a list of specific risk indicators that will increase or decrease the risk of readmission for a selected patient. The tool also provided specific evidenced-based recommendations to address the most significant risk factors. Over 200 care managers have use these insights in real-time to select and implement personalized care interventions (e.g., home health, remote monitoring, pharmacist consult, etc.) to mitigate a patient’s specific risks on more than 300,000 discharges over 3 years.

Conclusions
Insights from predictive models for readmission risk can be used in real-time at the point of care to potentially reduce readmission rates by customizing the discharge interventions to the risk elements that are germane to a specific patient.

Acknowledgements
This project was funded by Carolinas HealthCare System.

References
Integration of Genomic, Specimen, and Cancer Registry Data in an i2b2 Data Analytics Platform for Oncology Research

Jack W. London, PhD¹, John Reber¹, Chirayu Goswami, MS¹, Robert Stapp, DO¹, Stephen Peiper, MD¹
¹Thomas Jefferson University, Philadelphia, PA,

Introduction

Precision Medicine research requires access to comprehensive data describing a patient’s clinical attributes, including their diagnoses, treatment and outcomes. Of particular value are any molecular diagnostic test results noting genomic mutation expressions, or lack thereof. For oncology research, the basic patient diagnostic and treatment information found in the EMR must be supplemented by more comprehensive information, such as disease stage, recurrence, treatment details, and patient outcomes. These data can be obtained from the institution’s cancer (or tumor) registry. Additionally, investigators often need information on available well annotated, properly acquired and banked biospecimens for their research, as well as ready access to slide images. The Sidney Kimmel Cancer Center at Thomas Jefferson University has developed a central comprehensive data resource to provide investigators with these data.

Methods

The i2b2 platform was used to model patient data concepts for data obtained from Jefferson hospital’s EMR, cancer registry, biobanks, and pathology clinical systems. Unique ontologies (taxonomies) were developed to describe the concepts relating to cancer registry, biospecimen, and molecular diagnostic genomic test results, and integrated with the diagnostic, procedure, medication, and laboratory concepts for patient data obtained from Jefferson’s EMR. The data in the resulting i2b2 data mart are refreshed on a weekly basis.

Results

An i2b2 research data mart describing over 1 million patients with ~100 million observations, including molecular diagnostic results on over 300 genes and 3,000 associated mutations is available to researchers. Detailed oncology data including disease stage, histology, treatment, recurrence, and survival data on over 100,000 patients from Jefferson’s cancer registry are incorporated. The i2b2 data is de-identified, but may be re-identified when necessary and with IRB approval. An honest broker system is used to perform tasks such as biospecimen acquisition. The i2b2 query tool is used for cohort identification, and a data visualization tool from TriNetX Corporation (Cambridge, MA) is used to explore cohort characteristics for clinical design and hypothesis generation. Links from the specimen banking system, OpenSpecimen, provide direct access from the i2b2 query tool to a web-based pathology image viewing application.

Conclusion

Identification of patient cohorts for precision medicine clinical trial research is facilitated by this comprehensive data resource. This is not only for patient recruitment, but also for clinical trial design. By determining the size of recent patient populations corresponding to proposed clinical trial eligibility criteria, the number of trials failing to accrue sufficient patients can be reduced. These comprehensive integrated clinical, genomic, cancer, and specimen data resources also facilitate hypothesis generation by defining associated characteristics of patient populations.

References

Multiword Frequency Analysis Based on the MEDLINE N-gram Set

Chris J. Lu, Ph.D.¹, ², Destinee Tormey¹, Lynn McCreedy, Ph.D.¹ and Allen C. Browne¹
¹National Library of Medicine, Bethesda, MD ²Medical Science & Computing, LLC, Rockville, MD

Abstract

Multiwords are vital to better precision and recall in NLP applications. The Lexical Systems Group (LSG) developed an effective approach to add multiwords to the SPECIALIST Lexicon from the MEDLINE n-gram set. This paper describes a frequency analysis on LexMultiwords (LMWs) and acronym expansions (e.g. blood pressure for BP) based on the word count (WC) in MEDLINE. Results show most LMWs locate in the low WC range with better precision and F1 score.

Introduction

LMWs are terms in Lexical records containing space(s). To be in the Lexicon, these terms must: 1) have a single part-of-speech (POS), 2) have inflections, and 3) be a special unit of lexical meaning by themselves. A set of filters and matchers based on empirical models has been developed to retrieve LMW candidates from the MEDLINE n-gram set¹. This process generates high precision LMW candidates for efficient LMW building. An analysis of WC allows us to use the frequency filter effectively for better LMW acquisition.

Approach

First, all unique single words (464,781) and multiwords (431,432) in Lexicon.2015 are retrieved. Second, the acronym expansions are retrieved by applying the Acronym Expansion Pattern (AEP) matcher to the MEDLINE n-gram set. This AEP set includes 14,440 LMW candidates. They are tagged by LSG linguists and are added to the Lexicon if they are valid LMWs. The WC from the MEDLINE n-gram set is added to these three data sets to derive the frequency spectrum of WC class vs term number. Term number (TN) is the total terms in a WC class with a range of 100 incremental (Figure 1). The frequency spectrum of WC class vs local precision (valid tags/total tags), recall (valid tags/total valid tags), and F1 measurement (PRF) are derived for AEP (Figure 2). The recall is normalized between 0 and 1.

Conclusion

Figure 1 shows that most LMWs are located in the low WC range for both the Lexicon and the AEP. This result coincides with the distribution of single words in the Lexicon and “Alice in Wonderland”². It seems both single words and multiwords share the common characteristic of distributing in the low frequency range. Figure 2 shows that low frequency n-grams in the AEP have higher normalized recall and F1 score, with precision above 0.8. Very few LMWs exist in the high WC range with a variation in precision of either 0 or 1. Accordingly, the frequency of LMW acquisition should be set on the lower WC range (100-10k) while the frequency of single word acquisition is set on the high WC range (because most unigrams are valid single words). This frequency strategy is applied with filters and matchers to generate LMW candidates from the MEDLINE n-gram set to enrich the coverage of LexMultiwords. Ultimately, this enhanced coverage provides better NLP results for projects that use the SPECIALIST Lexicon, and our WC results may guide MW acquisition efforts in others’ datasets.

References

How Strong is the Evidence for the Effectiveness of Dementia Family Caregiver Information and Communication Technology Interventions?

Robert J. Lucero PhD, MPH, RN, FAAN1,2, Aditi G. M. Patel, MPH, BS3, Elizabeth Fehlberg, BSN, BA, RN1, Michelle Santoni-Miranda, BA1, Karis Lee1, Renessa Williams1

1College of Nursing and 3Medicine, University of Florida and VA HSR&D Center of Innovation on Disability and Rehabilitation Research, Gainesville, FL, USA

Introduction

Information and communication technology (ICT) enabled interventions can improve decision confidence and self-efficacy, decrease spousal relationship conflict and activity restriction, and reduce emotional strain and burden in caregivers of persons living with dementia1,2. The Internet has been used to make available disease-specific information, private email, question-and-answer forums, monitoring and counseling, in-home support groups, and telephone access to information about caregiver issues3,4. The purpose of our study was to evaluate the current state of empirical research about the effects of ICT interventions on the health of dementia family caregivers (DFC).

Study Design

We conducted a systematic review. We searched PubMed, CINAHL, PsycINFO, and the Web of Science for literature referencing ICT, dementia, and caregiving concepts without date limits. We included randomized controlled trials (RCT) conducted in the United States in English or Spanish with non-professional caregivers. Non-RCTs, studies not in English or Spanish, and studies of professional caregivers were excluded. We used the Effective Public Health Practice Project (EPHPP) Quality Assessment Tool for Quantitative Studies in our review.

Population Studied

Family caregivers of non-institutionalized family members diagnosed with dementia.

Principal Findings

We found 523 articles and removed 21 duplicates. An additional 13 articles were identified from the reference lists of 2 systematic reviews. Two primary reviewers screened the abstracts, which resulted in removing 203 articles. A full-text screening by two secondary reviewers resulted in taking out 300 articles. The secondary reviewers conducted the quality assessment of 12 articles. Their independent ratings across six assessment components (i.e., selection bias, study design, confounders, blinding, data collection method, and withdrawals and dropouts) yielded an interrater agreement of 83.33%. Selection bias was rated as weak in a majority of the studies (i.e., 58%). However, 75% to 100% of the studies received moderate to strong ratings for all other quality assessment components. The secondary reviewers met with the principal investigator to clarify differences in ratings. Based on the EPHPP guidelines, this review found 4 strong studies, 5 moderate studies, and 3 weak studies.

Conclusions

These early findings provide a glimpse at the current state of the science on the effectiveness of ICT interventions with DFCs. These studies provide a viable foundation for ongoing discovery of effective ICT solutions. As consumers of health care adopt ICTs, it is important that ICT-enabled interventions improve health, including decreasing activity restriction and reducing emotional strain and burden in family caregivers.

References


Acknowledgment: New York City Hispanic Dementia Caregiver Research Program. NIH R01NR014430.
A Research Capability Framework

Airong Luo, PhD1, Marcelline R. Harris, PhD, RN2, Frank J. Manion, MS3, Barbara Mirel, D.Arts4

1Research IT Med School, 2School of Nursing, 3Cancer Center Informatics, 4School of Education, University of Michigan, Ann Arbor MI

Research capability is the competency to leverage resources and services for purposes defined by the goals of the research project. The rapid maturing of research informatics as a sub-discipline within biomedical informatics reflects, in part, a growing awareness of the need for new types of sustainable research environments that support multi-scalar and multi-disciplinary collaborative science. While a growing literature addresses capability models (e.g., software capability maturity1, scientific data management capability maturity2), and features of consortia and collaboratories3,4, no framework systematically assesses researchers’ needs for informatics support, and assists in institutions’ strategic planning. In addition, research capability framework is needed to provide a bridge between the scientific community and solution-focused research supporting units and organizational leaders.

Methods We interviewed 40 biomedical researchers in different fields at an institution characterized as “very high research activity”. We complemented the interviews with an in-depth case study of a highly productive research group that exhibited collaborations at various scales5. Next, 130 researchers within a comprehensive cancer center responded to a survey to identify needs and gaps in research informatics. We also reviewed the literature to identify factors leading to successful research performance in consortia and collaboratories. Interview data were analyzed using standard qualitative data analyses techniques (AL & FM), the case study was analyzed using visualization techniques (BM), and the literature was analyzed using integrative synthesis techniques (MH). We jointly iterated across these analyses to identify factors that enable research capability across a range of research requirements and collaborations. Based on our data analyses, we developed RCF, which consists of six modules of questionnaire.

Results One RCF module gathers profile information; the other five modules focus on components of scientists’ research processes empirically derived from our data. They include: “collecting, storing and transferring data,” “discovering, accessing and retrieving data,” “defining data to facilitate sharing, linking and discoverability,” “interpreting data through statistics and bioinformatics,” and “addressing governance, legal and commercial issues.” Each research process module consists of three “layers” of capacities and capabilities that were identified from our data as essential to achieving the goals of the research teams: people and organizational layer, information layer, and a technology layer. For each set of questions scientists specify how well supported or sustained they and their collaborators are in conducting research activities on a scale of 1-5. They also rate whether they are satisfied or not, and – if dissatisfied - identify the support that would satisfy them more. Given the user-centeredness of the survey, responses reveal the extent to which sub-processes in the research workflow are systematically supported, that is, maturity levels of the various technological and organizational support. Responses reveal patterns/groupings in practice across different profiles of biomedical researchers, which, in turn, can show successes and gaps in the support these groups respectively need. Validation suggests that RCF captures key resource and service requirements across a range of types of research and collaborations.

Discussion We are ready to administer RCF to a large group of diverse biomedical researchers. RCF has great potential to help researchers specify their needs, requirements, and priorities for centralized support, resources, and for funding proposals, thus enabling the specialists responsible for providing support to identify users’ needs and requirements in the context of research workflows, leading to appropriate solutions.

References
Using Results of Statistical Text Mining in Big Data Analysis

Stephen L. Luther, PhD, James McCart, PhD, Dezon K. Finch, PhD, Lina Bouayad, PhD, William Lapcevic MS, MPH, Michael Matheny, MD, MS
HSR&D Center of Innovation on Disability and Rehabilitation Research.
James A, Haley Veterans Hospital, Tampa, FL

Introduction
The Veterans Health Administration (VHA) has one of the largest electronic health record (EHR) systems in the world including more than 3 billion text records. The VHA has invested significant resources to develop infrastructure to extract and analyze text-based data. An ongoing study aims to improved risk models for pressure ulcers in Veterans with spinal cord injuries (SCI). While ICD-9-CM codes are available for documentation of pressure ulcers, under-coding is suspected which would impact the validity of the dependent variable for the analysis. Here we discuss efforts to use statistical text mining (STM) to find the occurrence of pressure ulcers.

Methods
All available data from FY 2009-2013 were obtained from a cohort of 11,884 Veterans with SCI seen in VHA SCI Centers during FY 2009 who had no recorded pressure ulcers in the previous year. This resulted in approximately 13 million text documents and several million inpatient and outpatient administrative records. A sample of 2,500 documents were chosen from across VHA facilities and independently reviewed by two clinical annotators and adjudicated by a clinical expert. The data were split into a training (n = 1,750, 70%) and test set (n = 750, 30%). STM models employing support vector machines were built and compared via 5x10 fold stratified cross validation on the training set to select model parameter values. Using the best performing parameters, a final model was built from the training set and achieved sensitivity = .85, specificity = .91, positive predictive value = .87 and an F measure = 0.86 on the test set. This model was then applied to all 13 million documents in the cohort. In parallel, inpatient and outpatient records for the cohort were searched to identify outpatient visits and inpatient admissions in which at least one ICD-9-CM in the range of 707.00 to 707.9 was included.

Results
There were 3,336 patients (33%) classified as having a pressure ulcer based on ICD-9-CM codes. There were 10,218 patients (86%) classified as having a pressure ulcer based on the STM results (at least one document with a confidence of .50 or higher). One strategy to reduce the likelihood of false positives in this analysis would be to interpret higher cut points. Table 1 describes results of setting higher cut points in our test sample (n = 750). As expected, there is a sensitivity/specificity trade off, with little room for improvement in specificity and PPV beyond a cut point of .70. Another strategy to reduce false positive reduce would be to summarize positive documents over time. Analysis of the thirty day window after the first positive pressure ulcer document found that many Veterans had multiple positive documents. A combination of higher cut points and patterns of positive documents across time will be used to explore statistical methods to maximize information from text in this study.

Table 1. Effect of Varying Cut Point on Results from Test Sample

<table>
<thead>
<tr>
<th>Cut point</th>
<th>Accuracy</th>
<th>F-measure</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.8840</td>
<td>0.8508</td>
<td>0.8239</td>
<td>0.9243</td>
<td>0.8794</td>
<td>0.8868</td>
</tr>
<tr>
<td>0.6</td>
<td>0.8387</td>
<td>0.7604</td>
<td>0.6379</td>
<td>0.9733</td>
<td>0.9412</td>
<td>0.8004</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7573</td>
<td>0.5708</td>
<td>0.4020</td>
<td>0.9955</td>
<td>0.9837</td>
<td>0.7129</td>
</tr>
</tbody>
</table>

Conclusion
The cut point of .50 for confidence for a single document is commonly used to interpret results of STM models at the patient level. In this study the simple interpretation of this convention would suggest that nearly 90% of the cohort had a pressure ulcer. While this important clinical outcome is likely under coded in administrative data, this three fold increase over results using ICD-9-CM alone seems unlikely. However, when using STM to summarize multiple documents across time to classify patient level attributes more sophisticated techniques might be preferred. Here we describe a strategy to combine of cut points of predicted probabilities and patterns of positive documents to interpret data from a large observational cohort study in the VHA.
Use of SNOMED CT Relationships for Optimizing Automatically Extracted Problem List Entities from Clinical Narratives

Sina Madani, MD PhD1, Jerry Henderson, MD2
1Department of Institutional Analytics & Informatics; 2Department of Emergency Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX

Introduction
Natural Language Processing (NLP) solutions, such as MetaMap, can be used for extraction of medical concepts from clinical narratives. MetaMap uses an open architecture to perform lexical and syntactic analysis of the input text and generate structured data such as problem list. The patient problem list in a structured format with backend mappings to standard terminologies like SNOMED CT is an essential component of the electronic medical record (EMR) systems. However, problem list entities extracted with NLP methods may contain similar concepts with various levels of granularities. These concepts that are often related to each other may seem duplicate from physician’s perspective. During the transition period from our legacy system to Epic, we investigated whether we can eliminate such duplicate concepts in the problem list by using SNOMED CT ontological relationships.

Methods
MD Anderson Cancer Center’s EMR repository contains more than 15 million transcribed notes. For our study, we selected the last ten instances of the clinical notes for all patients. We defined a patient “Problem” as something that requires a plan for diagnosis and/or management. Based on this definition, we only analyzed part of the clinical notes that pertained to “Assessment & Plan” section headers. We limited MetaMap output to the Disorder semantic group and an extended version of the UMLS Clinical Observations Recording and Encoding (CORE) Problem List subset of SNOMED CT (v2015). Two reference tables were generated from a Web Ontology Language version of SNOMED with pairs of concepts that have “is-a” or “attribute” relationship. These relationships in SNOMED CT connect concepts in the same or different hierarchies. We eliminated the most general concept in such relationships and kept the most granular one only if all concepts seen together in the same problem list. Subsequently, we analyzed the effect of such reduction in the number of the reported problem list entities for all patients.

Results
We identified 2,966,344 clinical notes from 496,901 patients who had two or more clinical problems. The broader concepts like “Renal Mass” in an is-a relationship pair (Renal Mass - Angiomyolipoma of Kidney) or “Small Cell Carcinoma” in an attribute relationship pair (Small Cell Carcinoma - Small Cell Lung Cancer) were eliminated from patient problem lists during NLP post processing step. More than 80% of the problem lists had a reduction between 5% to 25% in the number of the reported problem entities (Figure 1). The average number of the dropped concepts from the problem list was 1.65 in which the largest drop (7 concepts) was observed in the 55% reduction bin.

Discussion
This study shows that similar clinical problems with various degrees of granularities can be detected from patient charts and adjusted by using SNOMED CT relationships. Use of the most detailed format of a clinical concept in the patient problem list may result in a more accurate documentation and better mappings to billing codes with higher granularities. Further analysis is still needed for identification of other sources of concept duplications caused by NLP output. Such analysis is particularly important in NLP methods that are often dependent on external contexts for disambiguation purposes, while a pre-coordinated context, such as “history of - “ or “other –“, has already been embedded internally in the clinical concept itself.
Reducing Clinician EHR Screen Time and Increasing Patient Face Time

J Edward Maddela, MD1, Khaleel Hussaini, PhD2, Umar Iqbal, MD3
1,2,3University of Arizona College of Medicine, Phoenix, Arizona

Introduction

EHR optimization entails reducing clinician EHR “screen time” and increasing patient-provider “face time.” Though early adopters have already begun this work1, few studies in peer-reviewed journals provide evidence-based guidance2,3,4,5. Consequently, enhancing EHR use post-implementation remains a daunting challenge for applied informaticians. Should optimization efforts begin with improving clinicians’ documentation time, computerized provider order entry (CPOE), or chart review time? This study aims to answer this question through a quantitative, data-driven method to optimize EHR efficiency for ambulatory providers across three medical specialties within a large hospital system.

Methods

We obtained three months of de-identified, aggregated, timestamped EHR data for providers (physicians, nurse practitioners, and physician assistants) employed by a large health care system within the medical specialties of Family Medicine, Internal Medicine, and Pediatrics. A total 138 providers were included in the final analytic sample for whom complete data were available. All providers used the same EHR. We utilized timestamped activity log data captured in the EHR using a response time measurement system. The primary outcome variable of interest to measure efficiency was Actual Time per Patient, defined as “the active time a provider spends within the EHR for each patient.” Univariate, bivariate, and multivariable statistical procedures were conducted using SAS v9.4 (SAS Institute, Cary, NC). Where appropriate variables were log transformed to meet Gaussian distributional assumptions of linearity, independence, and homoscedasticity. We tested the assumption of equality of variances after transforming of our primary outcome variable of interest using Levene’s test for homogeneity of variances and then performed one-way Bonferroni corrected ANOVA to test means between the three specialties to adjust for multiple comparisons.

Results

Internal Medicine providers spend more time in the EHR (27.1 minutes) than Family Medicine (22.9 min) and Pediatrics (10.4 min). Summary page chart review, documentation time, and flowsheet time explained over 90 percent of the variance in actual time per patient in the EHR. Irrespective of specialty, providers who spent more than five minutes on the summary page were 70 times more likely to spend over 22 minutes on Actual Time per Patient in the EHR.

Conclusion

The main goal of this study is to show how EHR usage data can be leveraged to optimize provider efficiency. Based on the data, CPOE does not influence the Actual Time Spent per Patient as strongly as chart review (summary page and flowsheet) and documentation. Thus, helping providers optimize these specific areas may produce a greater impact on their EHR efficiency. Despite its limitations, similar statistical modeling can be applied to other medical specialties, but more research is needed. Ultimately, reducing provider “EHR screen time” can potentially enhance clinical efficiency, improve provider satisfaction, and increase much needed quality “face time” with patients.

References

In Patient Usability Testing Results of a Patient-Centered Postoperative Wound Surveillance Smart Phone App Utilizing Digital Photographs
Andrea N. Mahnke, MS1; Rebecca Gunther, MD2; Sara Fernandes-Taylor, PhD2; Kate L. Baker, MA1; Jason T. Wiseman, MD2; K. Craig Kent, MD2
1Marshfield Clinic Research Foundation Biomedical Informatics Research Center, Marshfield, WI
2Department of Surgery, University of Wisconsin-Madison, Madison, WI

Background
Surgical site infection (SSI) is the most common nosocomial infection in surgical patients and accounts for 38% of post-operative complications1. Patients rarely recognize early stage wound infections and often require intensive treatment and rehospitalization2. It also produces increased health care costs and is the leading cause of unplanned, potentially preventable hospital readmissions for surgical patients1. The fact that SSI develops or progresses in the outpatient setting makes transitional care coordination an important focus in the effective management of SSI. Patient-centered technological interventions to improve transitional care for surgical patients are necessary to stem the burden of SSI and readmissions.

Objective
We designed and tested a patient-centered, outpatient wound surveillance program using smartphone digital photography to promote early recognition of SSI following discharge with the goals of (1) empowering patients to partner with their surgeons in monitoring their SSI symptoms and overall postoperative recovery, (2) diagnosing SSI at an early stage, enabling outpatient management, and (3) preventing hospital readmission and the serious morbidity and mortality associated with wound complications.

Methods
We designed a patient-centered smartphone app that enables patients to transmit wound photos and symptom information to the vascular service. In collaboration with geriatricians and community/patient research advisors, we designed a patient-centered training process to teach patients or caregivers to take and transmit photos of their wound using smartphones. We undertook formal usability testing of the app on inpatient surgery patients to evaluate (1) patient experience, (2) picture and information quality, and (3) successful information transmission and assimilation into PACScan. Screen and audio recordings were captured during the inpatient usability testing for analysis and identification of usability issues. Improvements were recommended and implemented iteratively.

Results & Conclusions
Two rounds of testing were conducted with a mix of vascular and oncology surgery patients. Round 1 (n=4) on Nov. 4, 2015 and Round 2 (n=5) on Dec. 11, 2015. Testing was done with a mix of race, sex, age and smart phone familiarity. Caregivers assisted with three of the nine participants. Smart phone training was given to patients who were not already confident working with a smart phone, but all participants received a brief training on the app. Only two Round 1 participants required guided app training and practice, the rest were confident to try the app on their own after the initial brief training. Significant usability issues were found after Round 1 testing and improvements were implemented for Round 2. All participants were able to successfully complete stepping through the app with their wound with no or some assistance, but none failed to complete. The average System Usability Scale (SUS) score for Round 1 was 85 and 82 for Round 2. Broadly interpreted, scores under 60 represent systems with poor usability; scores over 80 would be considered above average.

Supported by the Agency for Healthcare Research and Quality R21 1R21HS023395 R21 HS023395

References
Proactive Patient Management Worksheets: an opportunity to support clinical care while addressing increasingly complex reporting requirements

Julie Maitland, PhD¹, Natalia Largaespada Beer, MD², John N. Rutter, BSc Eng¹, Jeff Beairsto BSc Eng¹, Iain MacNeil BSc¹, Colin Kilburn BSc¹, Ian Smith, BSc²
¹Populus Global Solutions Inc., Fredericton, Canada; ²Ministry of Health, Belmopan, Belize

Introduction

When considering the reporting burden of developing countries, Belize is at an advantage over many countries thanks to the Belizean Health Information System (BHIS), an integrated HIS that captures structured data from clinical encounters within the public health care system. In addition to the clinical and operational reports embedded within the BHIS, the Belizean Ministry of Health (MOH) has a skilled IT team who work with the Epidemiology Unit and various Ministry Technical Advisors (TAs) to extract and report on custom queries. Despite this, the burden of timely reporting is not an insignificant one; one which is only going to increase with the advent of the United Nations’ Sustainable Development Goals¹.

Approach

In 2015 we embarked on a project with the Belizean Maternal and Child Health (MCH) Program. The original goal of the project was to increase the reporting capacity and autonomy of the MCH team. Over the course of the project, the focus shifted from after the fact reporting to the design of proactive patient management through BHIS-integrated worksheets; creating a secondary goal of providing access to reportable data in a format that could facilitate proactive patient management, helping to identify and manage the cases that are at risk of slipping through the net and becoming an undesirable statistic identified in after-the-fact reports. With a focus on Antenatal Care and Child Immunization, we carried out User-Centered Design activities with Ministry of Health Personnel, Regional Managers, and Clinicians “on the ground” to ensure that the worksheets were configured to reflect existing protocols and designed to provide value to as many MCH stakeholders as possible.

Worksheets

The worksheets, that are automatically populated by data already being collected in the BHIS during clinical encounters, list the details of a specific cohort of patients that a clinician or administrator can work from. Each worksheet provides an overview of how well each patient’s care adheres to the appropriate clinical protocol. Each entry links to the patient’s e-health record where data can be reviewed in more detail and new encounters can be documented. The Antenatal Care Worksheet provides an overview of women who are currently or were recently pregnant, listing all clinical encounters, last recorded measures, diagnostic screening status, and pregnancy outcomes. The Child Immunization Worksheet provides an overview of children and their immunization status and child growth measures. It is hoped that the worksheets will enable health professionals to identify and manage high risk patient cases in real-time, e.g.: by enabling a Community Health Nurse to generate a list of infants who are currently due/overdue vaccines in a particular district, or by enabling a Nurse Manager to generate a list of women who have not yet received a postnatal visit.

Conclusion

We suggest that protocol-based patient management worksheets can inform and enhance clinical practice while increasing the accessibility of the data required by international reporting and funding obligations. Both worksheets are in the process of being deployed at the time of poster submission. The success of the project will be measured in terms of the MCH team’s reporting capacity and autonomy, their use of the worksheets to inform clinical practice, and the identification, management, and health outcomes of high-risk mothers, newborns, and children.

References

Incorporation of Case Setup Complexities into the Analysis of Operating Room Turnover Times

John D. Manning, MD¹, Manmeet Singh, DO¹, Pavithra I. Dissanayake, DO¹, Elizabeth Day¹, MPH, Richard J. Banchs, MD¹

¹The University of Illinois at Chicago, College of Medicine, Chicago, IL

Introduction
The operating room (OR) achieves optimal efficiency when it provides the highest quality of care with the lowest use of time, money, and space [1]. Parameters such as case duration, OR workload, and case turnover times (TOTs) may be automatically generated for analysis [2], though factors such as TOTs are subject to wide variation [3]. Since the setup for a hernia repair differs vastly from that of a spinal fusion, we sought to examine if separating cases by complexity would impact their average TOTs.

Methods
This was a retrospective study at a single academic tertiary hospital. Case setup complexity levels were determined a priori using a panel of physicians and surgical nursing staff. All OR cases for a single month (December 2015) were placed into one of three category tiers. TOT calculations were then made using Compass Portal (The Advisory Board Company), with customizations to view each OR case individually. Data was analyzed using Minitab Statistical Software.

Results
A total of 1,177 cases were observed, of which 768 were eligible for TOT calculations in Compass Portal. To avoid outliers (also called delays), cases with TOT of more than 90 minutes were excluded (n=98). Of the remaining 670 cases, 173 were labeled Type 1 (simple, 25.8%), 348 were labeled Type 2 (intermediate, 51.9%), and 149 were labeled Type 3 (complex, 22.2%). Mean TOTs and standard deviations for each case are listed above [Figure].

Conclusions
We demonstrate significant variation in TOT at a single institution when separating cases by setup complexity. This may provide added benefit in TOT workflow assessments, such as focusing on TOTs for less complex cases. Improved accuracy of case complexity TOTs may also improve the accuracy and efficiency of OR scheduling. The potential also exists to identify surrogate markers of setup complexity, such as the number of trays opened per case. Lack of objectivity in determining setup complexity is a possible limitation of our study. Nevertheless, we believe that our classification system offers utility for optimizing OR TOT and can be readily adapted to other institutions.

References
Identifying Treatment Refusal Through Natural Language Processing
Joel J. Martin, MS, Nathaniel S. Ring, MA, Andrea A. Kelley, MPH, Jon D. Duke, MD
Regenstrief Institute, Indianapolis, IN

Introduction
Patient refusal of treatment (including medications and procedures) is a valuable data point for research and clinical quality initiatives but is poorly captured in traditional structured data sources. Natural language processing (NLP) has the potential for increased yield of information related to treatment refusal but has not been well studied for this purpose. Previous research in this area has been limited to colonoscopy refusal. No studies have addressed identification of general treatment refusal events. In the current project, we sought to examine NLP performance in identifying treatment refusal.

Methods
Our source data was the Indiana Network for Patient Care, a health information exchange with clinical data on thirteen million patients. Specifically, we looked at a subset of 2.8 million patients with unstructured documents available in the Regenstrief natural language processing platform. We developed the treatment refusal algorithm through an iterative process comprising the following steps: 1) heuristic token selection (e.g., refused); 2) generating synonyms and lexical variants; 3) term discovery from documents with structured refusal data (e.g., vaccination); 4) manual annotation; and 5) evaluation and iterative review.

We then ran the final refusal algorithm across a dataset of 95 million clinical documents and found 908,992 matches. A convenience sample of 400 of these documents was reviewed by four reviewers for the presence or absence of treatment refusal. To perform a comparison between capture of refusal information in structured and unstructured data, we paired our refusal algorithm with a set of vaccination terms (e.g., pneumovax, MMR) and identified 4,603 matching reports. Again, we selected 400 random documents from this set for manual review. In this case, reviewers recorded whether the patient refused an offered vaccine.

Results
Our general refusal algorithm achieved a 83% PPV, with false positives primarily due to the use of the term ‘decline’ in the setting of laboratory values or failing health. Our algorithm looking specifically at vaccination refusal yielded a 78% PPV. When comparing of NLP-based vaccine refusal detection compared with structured vaccine refusal data in the same population show, 3440 unique patients were identified with vaccine refusal via NLP compared with 977 documented in structured data, and 85 patients that were identified in both.

Conclusion
Treatment refusal can be identified with good accuracy using unstructured data. For vaccination refusal, a 350% increase was observed using NLP compared with structured data. While this approach requires additional refinement, NLP holds promise for improving the capture of ‘non-events’ such as refusal in electronic medical records.

References
Comparison of Electronic Health Record Data Sources to a Gold Standard Patient Data Set in Correctly Identifying Chronic Conditions

Shelby J. Martin, MS, RD\(^1\), Nicole Weiskopf, PhD\(^1\), David A. Dorr, MD, MS\(^1,2\)
\(^1\)Oregon Health & Science University, Portland, OR, United States; \(^2\)General Internal Medicine & Geriatrics, OHSU, Portland, OR, United States

Introduction. Electronic Health Records (EHRs) may store data about a patient’s diagnosis in one or more areas depending on how the user interacts with the EHR, the workflow required for billing or documentation, and the perceived relevance of the diagnosis. Diagnoses may vary in accuracy and reliability based on the location of the diagnosis,\(^1\) potentially affecting key tasks related to prognosis, treatment planning, and prediction of adverse outcomes from disease burden. We sought to visually characterize the trade-offs between different EHR data sources of diagnoses.

Methods. We queried data about adult patients seen in primary care from the EHR for any evidence of 13 conditions. Diagnoses that were determined to be “positive” from a structured gold standard chart review were compared with frequency of positive diagnoses in four EHR data sources (problem list, outpatient encounters, medical history, a phenotype-based rule, or combined - the presence of the condition in any one of the four sources). Diagnostic characteristics of each source versus the gold standard were calculated, and EulerAPE version 3 software was used to create diagrams to visually represent the trade-offs from each source.

Results. 2,152 patient diagnoses in the gold standard data set were reviewed. Specificity was highest for the problem list (0.82), while sensitivity was highest for the combined data sources (0.95), followed by medical history (0.55). Figure 1 shows the trade-offs between sources. Green represents true positives, yellow missed diagnoses from the data source, and blue false positive diagnoses from the source. The Problem List shows a large number of missed diagnoses (42%), while the Combined had the largest true positives (67.4%) and false positives (23.7%).

Discussion. Choosing different EHR data sources for diagnosis definition altered the accuracy and completeness of the cohorts; visual representations of the data may help users understand the trade-offs between these diagnoses.

Figure 1. Diagnoses identified as positive as a proportion of all gold standard diagnoses

References

Review of Existing Mobile Apps to Support Symptom Management for Adults with Heart Failure Using the Mobile Application Rating Scale

Ruth M Masterson Creber, PhD, MSc, RN, Grenny Hiraldo, Meghan Reading, BSN, MPH, RN, Sarah J Iribarren, PhD, RN; Columbia University, New York, NY

Introduction
Heart failure is the fastest growing cardiovascular condition in the United States affecting 5.7 million Americans (1). Smartphone apps have the potential to support heart failure symptom monitoring and self-management. The purpose of this review was to identify and assess the functionalities of apps targeting patients with heart failure to support symptom monitoring and self-management.

Objective: The purpose of this review was to identify and assess the functionalities of patient-facing mobile apps targeted towards supporting patients with heart failure with symptom monitoring and self-management.

Methods
We searched 3 online mobile app stores using multiple terms and combinations (e.g., “heart failure”, “cardiology”, “heart failure and self-management”). Apps meeting inclusion criteria were evaluated using the Mobile Application Rating Scale (MARS) (2), IMS Institute for Healthcare Informatics functionality scores (3), and Heart Failure Society of America (HFSA) guidelines for non-pharmacologic management (4). Apps were downloaded and assessed independently by 2 reviewers, interclass correlations were calculated and consensus was met by discussion.

Results
Of 3,636 potentially relevant apps returned, 35 met inclusion criteria. Most apps were excluded because they were unrelated to heart failure, not in English or Spanish or were games. Interrater reliability was high (two-way mixed CA-ICC=0.93, 95% CI: 0.68-0.99). AskMD app had the highest average MARS total (4.9/5) and subscale scores followed by WebMD (4.4/5), which also had the highest functionality score (11/11). More than half of the apps (66%) had acceptable MARS scores (>3.0). Factoring both the MARS and IMS functionality scores (Figure 1), the highest performing apps included: AskMD, webMD, Heart Failure Health Storylines, Continuous Care Health App and Symple. Only one, Heart Failure Health Storylines, met all of the HFSA guidelines for non-pharmacologic management. Peer reviewed publications were identified for only three of the included apps.

Conclusions
Few apps had both a high MARS and functionality scores. No apps had high scores and met all of the HFSA guidelines. Therefore, involving patients in the design for further development of a heart failure tailored app and meeting criteria proposed in the guidelines by the HFSA is recommended.

References
Implementing Computerized Provider Order Entry Increased Rates of Duplicate Laboratory Testing in Inpatient and Outpatient Settings

Patrick C. Mathias, MD, PhD, Noah G. Hoffman, MD, PhD
University of Washington, Seattle, WA, USA

Introduction

While many studies of computerized provider order entry (CPOE) demonstrate decreases in overall medication prescribing error rates, CPOE implementation has been associated with increased duplicate medication errors1,2. This increase in errors has been attributed to many factors, including inadequate clinical decision support for reconciling duplicates, added effort required in cancelling or modifying orders, and the ability of multiple users to enter orders nearly simultaneously3. To date, medication orders have been a large focus of work examining CPOE’s effect on duplicates because of patient safety issues, but there are significant impacts from duplication of diagnostic testing as well. Duplicate laboratory testing can contribute to iatrogenic blood loss from repeated unnecessary draws and has downstream impacts in increasing costs from additional testing as well as labor in cancelling recognized duplicates. The aim of this study was to determine the impact of CPOE implementation on cancellation rates for duplicate laboratory tests within one health system.

Methods

Test cancellations and volumes over more than 4 years (May 2011-September 2015) were queried from a single laboratory information system (LIS, Sunquest, Tuscon, AZ) used throughout the duration of the study. Two laboratories share the LIS and serve a tertiary care academic medical center (Site 1) and a county hospital (Site 2). Site 1 implemented CPOE in PowerChart (Cerner, Kansas City, MO) in May 2012, followed by implementation with the same EHR at Site 2 in September 2012. In May 2014, all of the health system’s outpatient clinics (which send testing to both sites) were transitioned to Epic (Verona, WI) with CPOE. Data included the dates of collection, ordering location, and a code signifying the reason for cancellation. Only cancellations for duplicate orders associated with an accession generated in the LIS were included (excluding automated cancellations). Monthly duplicate cancellation rates were calculated over 1 year periods before and after each CPOE implementation, and stratified by hospital and location type (inpatient or outpatient). Statistical comparisons of proportions of cancellations before and after each go-live were performed using a two-sided test for equality of proportions.

Results

Rates of cancellations for duplicate orders increased significantly after each CPOE implementation. The inpatient implementations were associated with increases in cancellation rate from 1.3 to 30 cancellations per 1000 accessions at Site 1 (p < 0.001) and from 1.9 to 18 cancellations per 1000 accessions at the Site 2 (p < 0.001). The system-wide transition in the outpatient EHR also increased cancellation rates at both sites, from 5.5 to 17 cancellations per 1000 accessions at Site 1 and from 11 to 29 cancellations per 1000 accessions at Site 2 (p < 0.001 for both sites).

Conclusions

Implementation of CPOE increased the proportion of duplicate laboratory orders requiring cancellation by ten-fold or more in the inpatient setting and by three-fold in the outpatient setting. These findings suggest that configuring CPOE to prevent the creation of duplicate laboratory orders should be considered during implementation to prevent the effort required to reconcile duplicate orders in downstream systems.

References

Ensuring Data Integrity from Anesthesia Information Management Systems: An Approach by the Multicenter Perioperative Outcomes Group

Michael R. Mathis, M.D.1, Sachin Kheterpal, M.D., M.B.A.1
1University of Michigan, Ann Arbor, Michigan

Introduction

Within the era of big data analytics, healthcare provider access to anesthesia information management systems (AIMS) has become increasingly available. However, with continued expansion, maintaining complete and accurate AIMS-derived data has emerged as a new challenge. Traditional AIMS data sources have included administrative data, a preoperative history & physical, and an intraoperative record. New sources of AIMS data have included nursing flowsheets, as well as blood bank, pharmacy, radiology, and pathology data. National and global registries have additionally allowed for linking patient comorbidities and outcomes to AIMS data. With this explosion of information accessible for guiding clinical assessment and decision-making, maintaining data quality has become paramount, and calls to ensure data integrity derived from AIMS have been made.1,2

Quality of Reporting Driving Quality of Care

Challenges to maintaining data integrity are numerous, and must be addressed if AIMS-derived databases are to be sustained as tools for outcomes research and quality improvement. Within the field of anesthesiology, adverse events are rare; to characterize such adverse events, AIMS data must necessarily be complete and accurate. If methods to ensure data integrity are not followed, patient, provider, and process measures associated with specific outcomes are likely to be inaccurate, and unable to drive quality improvement. Limitations of each data source used must be considered, such as the training of data entry personnel, method of entry, and data structure itself. Within our Multicenter Perioperative Outcomes Group (MPOG), we synthesize AIMS data and systematically address these pitfalls.

The MPOG Approach to Data Integrity

Within the MPOG research consortium, participating centers identify key personnel within their institution to maintain departmental and technical support. Interfaces are built to enable data transfer from each site to a central database. Next, a database “mapping” utility is used at each site to transform the structure of local AIMS data elements to a uniform centralized structure. Variable mapping, overseen by a clinician at each site, is performed across multiple data types, including patient, provider, and surgical characteristics.

Next, a “data diagnostics” process is used to ensure data validity. Across all mapped data, percentages of meaningful and complete data entry are noted and tracked chronologically. Accuracy of data mapped are recorded, and thresholds for success (with a mandatory minimum requirement) are publicly reported to all participating institutions. A “case-by-case validation” tool is next employed to determine if case times and caregiver management are feasible; feedback for re-mapping database elements can be given to each site. A “case viewer” enables caregivers to view centrally-uploaded AIMS data as it would appear in an intraoperative record, to provide further instruction on quality improvement.

With AIMS data mapped, validated, and uploaded to a central database, outcomes research and quality improvement projects are proposed and discussed at weekly MPOG conferences, using standardized manuscript and data query templates with instruction through an online curriculum. Accepted projects are subjected to research “data cleaning”, in which data queried are audited by investigators and further mapped to discrete concepts for research purposes. Numerical data are cleaned through establishing valid ranges, and employing monitoring artifact reduction algorithms.

Conclusions

Through the use of systematic measures to ensure data integrity, our MPOG consortium makes strides to improve data quality that have been previously greatly outpaced by data quantity. In developing rigorous data validation measures, high-integrity AIMS-derived data can be maintained, correlated to outcomes, and used to drive quality improvement.

References

MedicineMaps: A Tool for Mapping and Linking Evidence from Experimental and Clinical Trial Literature

Nicholas J. Matiasz, MS1,2, Wei-Ting Chen1,3, Alcino J. Silva, PhD2, William Hsu, PhD1
1Medical Imaging Informatics, Departments of Radiological Sciences and Bioengineering, University of California, Los Angeles, CA; 2Integrative Center for Learning and Memory, Departments of Neurobiology, Psychiatry, and Psychology, Brain Research Institute, University of California, Los Angeles, CA; 3Department of Computer Science, University of California, Los Angeles, CA

Introduction. A significant barrier in the translation of clinical research has been the inability to effectively explore the large information space of published experiments to catalyze hypothesis generation and validation studies. Researchers need to become familiar with a body of research of increasing size and complexity: in 2013 alone, over 700 papers related to neurofibromatosis type 1 (NF1) were indexed on PubMed. Achieving an integrated understanding of a disease requires a shareable, machine-readable representation that not only captures high-level (causal) relationships but also incorporates relevant supporting evidence about the studies. Having a formal way to represent causal connections using semi-automated graphical and interactive tools would provide the research community with a map of what is known, unknown, and disputed, thus facilitating experiment planning.

We present an open-source web application called MedicineMaps that facilitates clinical translational research by systematically and collaboratively capturing results of experimental studies and clinical trials reported in literature in a shareable, machine-readable way. We build upon the concept of research maps4-3 to formalize causal relations based on a taxonomy of clinical experiments and rules for integrating evidence from multiple studies.

Materials and methods. At its highest level of abstraction, our representation includes three main entities: INTERVENTION (a treatment administered to a patient suffering from a condition), OUTCOME (an operationalized measure to assess a patient’s response to the intervention), and RELATION (a directed relationship between an intervention and an outcome that describes how the outcome is expected to change in response to the intervention). Details of each experiment are captured as properties of each relation, and a SCORE provides a heuristic measure to convey the amount of evidence represented. Each score is based on the convergence and consistency of the relevant experiments.

Results. MedicineMaps is implemented as a web application that assists target users (i.e., clinicians and researchers) with extracting and mapping information from papers to generate medicine maps. Functionality of this interface includes: (i) structured forms for entering details about the study; and (ii) visualization of semantic predications as a collection of nodes and edges. During the annotation of individual papers, experiments with specific intervention–outcome pairs can be inputted using a structured form that includes a variety of relevant fields to capture each study’s context. MedicineMaps is developed with Node.js, a JavaScript-based runtime environment for creating web applications. Given that the information is encoded as graphs, we use Neo4j 2.2.1, a NoSQL graph database. The database is queried using the Cypher Query Language (CQL) of Neo4j. Medicine maps are visualized using Cytoscape.js.

References
Introduction

Well-structured controlled vocabularies such as MeSH are designed to address language problems encountered in information retrieval including polysemy (multiple meanings for the same term). In an analysis of the visibility of Game Theory in the biomedical literature, McCain reported an apparent case of inadequate polysemy control, based on a contextual co-descriptor mapping of the MeSH term GAMES, EXPERIMENTAL (Games designed to provide information on hypotheses, policies, procedure or strategies). GAMES, EXPERIMENTAL is subordinate (NT, Narrower Term) to GAME THEORY (a subject within MATHEMATICS--essentially the mathematical analysis of interactive decision-making involving players who anticipate other players’ decisions in making their own). By extension, GAMES, EXPERIMENTAL would refer to the decision scenarios (games) adopted by the players in an experimental setting. However, the contextual descriptor linkage map for GAMES, EXPERIMENTAL showed a frequent co-assignment of MeSH terms relating to nursing education and neuro-imaging as well as human behavior, cognition, and decision making. Here I report an analysis of the MeSH term assignment (indexing) patterns for terms co-assigned with GAMES, EXPERIMENTAL and the content of articles retrieved with this MeSH term. The goal is to demonstrate the effect of introduced polysemy—multiple meanings of “games” in experimental settings.

Methods

Contextual co-descriptor analysis of the top substantive MeSH terms co-occurring with each other and with GAMES, EXPERIMENTAL was used to map descriptor co-assignment patterns over two time periods separated for contrast: 1974-1999 and 2005-2015. A full-text content analysis of primary research reports indexed with GAMES, EXPERIMENTAL identified the “games” and game-related research goals reported in this literature.

Results

In the earlier dataset (1974-1999), TEACHING was one of two MeSH terms with the highest degree centrality in the co-descriptor network, linked to more specific terms in nursing education (the other was GAME THEORY, the immediate parent term of GAMES, EXPERIMENTAL). The retrieved literature focused primarily on the use of “games” in education of nursing students and hospital staff—the top “games” were Q&A quiz show/board game formats and simulation/role-playing activities. In 2005-2012, the MeSH term DECISION MAKING dominated the co-descriptor network. Removal of this term and remapping of the remainder highlighted two different aspects of decision-making. One dealt with interactive social behavior using game-theoretic scenarios (e.g., The Ultimatum Game) while the other focused on individual risky decision-making activities (e.g., the Iowa Gambling Game). fMRI was used to map brain activity during both individual and interactive decision-making. Primary research using “games” in biomedical education was still visible but a minor part of the literature retrieved.

Conclusion

If GAMES, EXPERIMENTAL is only and directly NT to GAME THEORY in MeSH, it should retrieve only articles reporting experimental work in game theory/economics. Here, however, two other definitions emerge: (1) everyday games/skills assessment involving two or more people and (2) single person solitaire games and gambling/risk-taking activities. This indexer-introduced polysemy will result in a high proportion of non-relevant documents, no matter which of the notions of a “game” in an experimental setting is relevant to someone searching on the term. Vocabulary designers, indexers, and searchers should be aware of the potential for indexer-introduced polysemy when term definitions are incomplete or overly broad in the context of their hierarchical relationships.

References

Extending the PCORnet Common Data Model to Encompass Emergency Department Clinical Research.

James C. McClay, MD, MS, Bret Gardner, BS
Department of Emergency Medicine, University of Nebraska Medical Center, Omaha, Nebraska

Background: In 2013 the Patient Centered Outcome Research Institute (PCORI) established the Patient Centered Outcomes Research Network (PCORnet.org). PCORnet contains 13 Clinical Data Research Networks (CDRN) based on hospitals and integrated delivery networks with full implementation of electronic health record systems (EHR). PCORnet established a robust organizational structure for sharing EHR derived data in support of comparative effectiveness research (CER) trials. One major initiative is development of the PCORnet Common Data Model, currently in its third version (CDM V3). Each participating institution in the member CDRNs maintains a standardized data mart based on the CDM V3. PCORnet participants have also entered into reciprocal IRB agreements and sharing of data. Fundamental to this data sharing is the ability to generate de-identified data sets based on CDM V3. The current CDM V3 was developed based on lessons learned developing prior research networks such as Mini-sentinel and does not generalize easily to other types of clinical research. Specifically, in its current iteration CDM V3 does not support many of the data elements relevant to Emergency Department (ED) based research as reflected in the recently released HL7 Data Elements for Emergency Departments (DEEDS) specification. We developed a set of pragmatic extensions to CDM V3 enabling data marts maintained by PCORnet CDRNs to participate in multicenter ED based research studies.

Methods: CDM V3 consists of 15 tables organized around a patient record ID. A patient may have multiple episodes of care. DEEDS consists of a set of data elements organized into process oriented categories organized around an episode of care in the emergency department. DEEDS contains a comprehensive set of data elements mapped to standardized codes such as LOINC and SNOMED. We mapped data elements categories from the DEEDS specification to the PCORnet CDM. Elements from both data specifications were removed from analysis if they were not directly related to patient care, were not relevant for the ED setting, or too detailed to be reliably found in institutional data marts. We determined missing concepts and attributes to extend the CDM tables in order to enable ED research. The ED based extensions to the CDM were modeled in a data mart.

Results: A subset of CDM V3 tables are relevant to clinical care (Figure 1). A reduced set of research oriented data element categories from DEEDS were then mapped onto the CDM. Changes to accommodate ED research include the following: addition of missing data elements to tables such as add other vital signs to VITAL; Adding further table information such as PROVIDER in order to capture other roles in the ED; adding values such as “once” to RX_FREQUENCY field in the PRESCRIBING table and adding “Admitted” to DISCHARGEDisposition in ENCOUNTER table.

Discussion: The PCORnet national research network represents a significant investment in infrastructure and organization in support of clinical research. We propose a set of extensions to the PCORnet CDM that could be implemented at any CDRN site to support ED base based research for CER, outcomes studies and clinical trials. These extensions include additions to value sets, new fields in tables and new tables. With the near universal implementation of ED specific data systems at all participating PCORnet CDRN sites the represents an unprecedented opportunity to advance emergency care based research with a minimal incremental investment.

References
An Analysis of the Utility of Coded Override Reasons for Drug-Drug Interaction Alerts at Eleven Sites

Dustin S. McEvoy, BS1, The DDI Team, Dean F. Sittig, PhD2, Adam Wright, PhD1,3,4
1Partners Healthcare, Information Systems, Wellesley, MA; 2University of Texas Health Science Center, Houston, TX; 3Brigham & Women’s Hospital, Boston, MA; 4Harvard Medical School, Boston, MA, USA

Introduction

Although potentially harmful drug-drug interactions (DDIs) have long represented a serious threat to patient safety, preventing them remains an unsolved problem in healthcare. Point of care alerts within computerized provider order entry systems have the potential to lessen this threat, but are often overridden which may limit their effectiveness. A previous study analyzing prescriber reasoning for overriding DDI alerts suggested that coded override reasons (ORRs) should be required on more DDI alerts to facilitate communication of override rationale to the pharmacy, thereby preventing delays in the prescribing process.1 Many institutions now offer coded ORR choices for their interruptive DDI alerts, but the utility of these ORRs has not been analyzed. We analyzed coded ORRs for DDI alerts in various electronic health record (EHR) systems to assess their coverage of potential reasons, clinical utility (in terms of the assessment of potential patient harm), and whether they might be used to improve alert specificity.

Methods

Coded ORRs for interruptive DDI alerts were collected from 7 different EHR systems used at 11 health care organizations across the United States. These ORRs were classified to determine the percentage of ORRs useful in the assessment of potential patient harm and in improving alert specificity. ORRs were classified as useful or not using the criteria established by Grizzle et al., which classified an ORR as useful to verification pharmacists if it documented that the prescriber was 1) aware of the drug interaction, and 2) had indicated measures to mitigate potential harm to the patient.1 To assess whether ORRs might be used to improve the specificity of DDI alerts, we added a third criteria, assessing whether ORRs contained information pertaining to when the alert should or should not fire. Results were analyzed to assess the usefulness of current ORRs offered for DDI alerts in EHRs.

Results

We collected 60 coded ORRs for DDI alerts from 11 health care organizations using 7 different EHRs. Of the 60 ORRs, 48 (80%) documented prescriber awareness of the potential drug interaction. However, only 28 (47%) of these ORRs also indicated that steps had been taken to mitigate patient risk. Example ORRs that did not document physician awareness of the DDI included “Not true allergy” (which is irrelevant to DDIs), and “Unverified”. These were found in systems where there was a single set of ORRs in use across multiple alert types. Examples of ORRs which failed to document steps to mitigate patient harm included “Physician approved” and “Benefit outweighs risk”. The utility of ORRs varied by institution, but only one institution had ORRs which were all considered to be useful by these two criteria. More than half (32) of ORRs contained information pertaining to alert specificity. Examples included “Patient tolerated before” (i.e., the alert should not have fired due to medication history), and “This alert is NOT useful" (i.e., the alert should not have fired because the interaction is not clinically significant).

Conclusion

Of the observed ORRs for DDI alerts, 20% failed to document prescriber awareness of the drug interaction. In systems using a single set of ORRs for all alert types, some ORRs were irrelevant to DDIs. While use of a single set of ORRs across alert types is more consistent, it creates situations where a provider could inadvertently choose an ORR unrelated to the alert. Less than half of the ORRs documented both prescriber awareness and steps to mitigate potential patient harm. As a result, institutions should consider reviewing the coded ORRs that they offer for DDI alerts to ensure that they clearly communicate prescriber reasoning, or removing them all-together to reduce strain on prescriber attention. Over half of DDI ORRs contained information that was potentially relevant to the alert logic, suggesting that ORRs may be a useful tool to collect information to improve upon DDI alert logic. However, they may be best used for a short period of time given that they do not appear to add significant clinical value.

References

RexDB & GSM Provide Iterative Cohort Selection for GWAS: A Case Study of Gains through Interoperability of Two Open-Source Technologies

Owen McGettrick1, Ezekiel Maier, PhD2, Natasha Sefcovic, PhD2, Leon Rozenblit, PhD, JD1, Paul Hodor, PhD2,3
1Prometheus Research, New Haven, CT, 2Booz Allen Hamilton, Rockville, MD, 3Seattle Children’s Research Institute, Seattle, WA

Introduction: The increasing availability of human whole genome sequences makes it possible to build large data resources for precision medicine (Stephens et al. 2015), allowing analyses such as genome-wide association studies (GWAS) on an unprecedented scale (Welter et al. 2014). We describe a computational infrastructure that solves two fundamental challenges in building such a resource: (1) the backend that stores and manages the data has to be scalable from thousands to millions and more of complete human genomes, without loss of performance; (2) the system should have an intuitive and configurable user interface for cohort selection, based on (potentially changing) metadata about subjects represented in the database.

Methods: The backend consisted of the previously developed Genomic Sequence Manager (GSM, http://tinyurl.com/BAH-AMIA2015), a novel system based on a Hadoop prototyping environment (Hodor et al. 2016), HBase, and Amazon Web Services (AWS). We created a user interface using RexDB, an open source platform for supporting research data management workflows. A novel RexDB component called RexGuide was used to configure linked filter screens to support incremental cohort selection. The demo application maintains a database of anonymous subjects initially populated from the 1000 Genomes Project (http://www.1000genomes.org) (1000 Genomes Project Consortium 2015), which includes basic demographic characteristics such as gender and population of origin. The data has been further enriched for demo purposes by several synthetic phenotypic groupings and genetic variant presence. RexGuide allows user to easily configure independent filter controls simply by specifying the data elements or expressions of interest. Once configured, RexGuide generates a user interface for incrementally narrowing the population set down to the specific cohort of interest. Interoperability between RexGuide and the GSM is through a simple HTTP protocol using an XML job configuration file and CSVs of the cohorts. A job, consisting of an experimental cohort and a control cohort, can be submitted for GWAS analysis on the GSM. RexGuide monitors the progress of the job and when GWAS results are returned users can view the raw results or an interactive Manhattan plot of the results set. Data or the Manhattan Plot can be exported for external use.

Results: The interface supports: population filtering through configured filter screens; cohort selection; cohort management (saving, searching, editing, etc.); GWAS job submission and monitoring; GWAS results listing, searching, sorting, and graphing; linking GWAS results automatically to SNP_DB. The panel below shows four screenshots for (1) selecting individuals, (2) filtering by phenotype, (3) graphing results, and (4) configuring auto-generated filter screens.

Conclusions: By integrating RexDB with GSM we have created a powerful and intuitive tool for GWAS analysis from large-scale data repositories. Encouraging is how easy it was to configure and modify the cohort selection interface through RexGuide, which provided a flexible mechanism for managing a variety of metadata. Use of AWS cloud services involved zero upfront infrastructure investment and has the potential of organic expansion of the system with future data growth.

References

Introduction: Nanomedicines can demonstrate improved targeting efficiency and reduced side effects over conventional formulations. Hence, there is a critical need to automatically synthesize knowledge and trends in nanotechnology research from an exponentially increasing body of literature. In nanomedicine, new engineered nanostructures and formulations are continuously being created, and Natural Language Processing approaches can semi-automate the cataloguing and tracking of the different nanomedicines being developed. The goal of this project is to automatically identify nanomedicines characteristics from the literature using name entity recognition. In this work, we present (1) the initial development the Engineered Nanomaterial Database [1] (END) to support the evaluation of nano-entity extraction; (2) the evaluation of current entity extraction systems developed for general English; and (3) the initial development of our own nano-entity extraction system (NanoB2B Entity Extractor).

Methods: The END nano-entity test data was developed using the Drugs@FDA Database which contains product inserts of a wide variety of FDA approved drugs. The data was manually annotated for entity mentions consisting of nanomedicine physico-chemical characterizations, exposure information, and biological response information of 42 FDA-approved nanomedicines. We evaluated the performance of two state-of-the-art named entity recognition systems developed for general English (StanfordNER [2] and OpenNLP NameFinder [3]). We compare the results of those systems to the initial development of our NanoB2B entity extraction system which uses lexical information about the entity as features into a supervised learning algorithm using the WEKA datamining package [4].

Results: We evaluated the StanfordNER, OpenNLP NameFinder and our NanoB2B Entity Extraction system using leave-one-out cross validation on the END nano-entity dataset. The results show that the state-of-the-art entity recognition systems developed for general English are not sufficient to extract the nanomedicine mentions with F-measures ranging from 0.1 to 0.7. However, our specialized NanoB2B entity extraction system produced initial results that show great promise; demonstrating an increase in F-measure for almost all entities with results ranging from 0.5 to 0.8.

Conclusion: Our results demonstrate that there is sufficient reason and promising proof to the validity of developing an entity extraction system specifically for nanomedicines. Our initial NanoB2B Entity Extractor which incorporating lexical information increases accuracy of the results for some nanomedicine entities. In the future, we will explore utilizing additional syntactic and semantic information as identified by MetaMap into our system.

Homeless patients driving innovation: Establishing a community innovation panel to create patient-facing technology in a large urban health clinic

D Keith McInnes, ScD1,2, Salem Johnston, BA2,3, Robert Hass, BA, BSE4
1Center for Healthcare Organization and Implementation Research, Edith Nourse Rogers Memorial VA Medical Center, Bedford, MA; 2Department of Health Law Policy and Management, Boston University School of Public Health, Boston, MA; 3Finance Office, Boston Health Care for the Homeless Program; 4Information Systems, Boston Health Care for the Homeless Program

Abstract
Homeless populations have complex medical needs and a diminished ability to address them. Information technologies such as cell phones and the internet allow ubiquitous communication and broad access to information, but are beneficial in a health setting insofar as they mesh with patients' needs and abilities. At Boston Health Care for the Homeless Program, we established a patient Community Innovation Panel to address the gap between homeless patients' low use of technology and the promise that technology holds.

Introduction
Homeless populations simultaneously exhibit complex medical needs and a diminished ability to address them. Cell phones and the internet allow ubiquitous communication and broad access to information, but are beneficial in a health setting only insofar as they mesh with patients' needs and abilities.

Description of the Patient Community Innovation Panel (CIP)
At Boston Health Care for the Homeless Program, we established a Community Innovation Panel to address the gap between homeless patients' low use of technology and the promise that technology holds. This panel, comprised largely of homeless patients, 1) Identifies problems facing our patient population and proposes technological systems to meet these needs; 2) Supervises a tech training program tailored to the priorities and abilities of Boston's homeless population; and 3) Recommends improvements to current technological systems from the perspective of homeless consumers of health services. The seed of this idea was planted last year when we created a Patient Portal to let patients request appointments and prescription refills – interest was high, but adoption was low. Patients helped us identify obstacles such as unfamiliarity with technology, fear of doing something wrong with a device, and shame at admitting one’s ignorance. Our initial response was to start a computer training program for patients, but this approach had limited capacity and was only partially meeting patients’ technology needs. To tackle the underlying problems more directly, we joined forces with a public health technologist to form a Community Innovation Panel of patients, researchers, and health workers. This panel includes 8 patients, with a variety of cultural backgrounds and a wide range of comfort with technology. The Panel has begun developing a multi-year technology strategy for the clinic, including use of text messaging to increase patient-clinic interaction; and addressing technology training needs with an innovative and scalable “train-the-trainer” approach that is creating a cadre of homeless peer technology tutors. This session will describe how elements of community based participatory research (CBPR) guided the concept and formation of the Community Innovation Panel, how homeless patients created and instituted the technology training curriculum and its peer tutoring model (including findings from training evaluations), and the functions that will be included in the automated text message system (ideas proposed by the panel include appointment reminders, adherence reminders for chronic conditions, weather alerts, and health-promoting advice).

Conclusions
The Community Innovation Panel shows great promise in making health information technologies more patient-centered at health care for the homeless programs and other safety net systems. This panel is developing a multi-year technology strategy for the homeless clinic, and addressing technology training needs with a “train-the-trainer” approach that will rely on homeless peers as technology tutors and trainers. This contributes to important goals of more patient-centered health care and increased patient engagement in their health.
Qualitative Assessment of Women’s Attitudes Towards a Tracking App for Phenotyping Endometriosis

Mollie M. McIlllop, MPH, MA¹, Tamer Seckin², MD, Noémie Elhadad, PhD¹
¹Columbia University, New York, NY; ²Endometriosis Foundation of America, New York, NY

Introduction

Endometriosis—a chronic disease of the female reproductive system in which the uterine lining grows outside the uterus—is associated with significant morbidity, a reduced quality of life (0.809 quality-adjusted life years/woman) and substantial health care costs and productivity loss [1]. It is estimated to affect 1 in 10 women [2]. Despite its impact and prevalence, little is known about the disease, and diagnosis is often delayed by several years because of the lack of phenotypic knowledge about it. Our long-term goal is to phenotype endometriosis—a particularly challenging task for such an enigmatic condition [3]. A first step is to elucidate the catalogue of signs and symptoms of the disease and their manifestations throughout a patient’s menstrual cycles. Creating such a catalogue is important for identifying clusters of symptoms and their interactions throughout a cycle that could in turn indicate potential disease subtypes. While there exist research surveys to collect patients’ experiences, they assume a static set of symptoms. In contrast, we aim to engage patients in becoming active participants in disease phenotyping through a mobile app called Phendo. Phendo will allow patients to track their endometriosis symptoms as they occur throughout their menstrual cycle, and allow its users to contribute tracking custom/individualized symptoms. To assess the need and elucidate functionalities for Phendo, we conducted a series of focus groups to explore patients’ attitudes towards a tracking mobile application for the sake of disease phenotyping.

Methods

A series of 5 focus groups were conducted with a total of 27 women from the New York metropolitan area over a period of three weeks. Women were recruited through emails to endometriosis patient advocacy groups and flyers posted on the Columbia University campus and through Twitter. Eligibility included official endometriosis diagnosis through laparoscopic surgery. Each focus group lasted approximately 90 minutes and a semi-structured focus group guide was used to direct the conversation. Each focus group was audio recorded and transcribed verbatim. Thematic analysis was carried out to identify major themes related to tracking and endometriosis.

Results and Discussion

Overall, focus group participants felt positive towards tracking, although some women worried that tracking would lead to being reminded that “nothing helps”. Most participants used some sort of tracking for one or more aspects of their endometriosis (e.g., pain, diet and GI issues, exercise, emotions), but they felt that existing tracking apps (e.g., period trackers, fitness, diet trackers) were not appropriate for the range of symptoms and granularity of symptoms they wanted to track. Several participants reported having created their own customized tracking system.

Women expressed a range of motivations to track: to better understand their own experience of the disease (i.e., discover and reflect on potential patterns of symptoms throughout their cycle); to explore strategies for symptom management (“what works and what doesn’t”); to help manage their work and personal schedule based on their past tracked data and forecasted flares; and to communicate better with their doctors. Participants expressed a strong altruistic will to contribute their data for the sake of research.

Our findings indicate that patients would welcome an app for phenotyping endometriosis that captures the wide range of endometriosis symptoms. The discovered themes also help elucidate different strategies to engage patients in tracking, which we are currently exploring in the design of Phendo. Contrary to the few existing apps that track menstruation and even endometriosis, Phendo will enable the research community to gather and analyze unprecedented phenotypic descriptions of the disease.

References

Qualitative Analysis of Proxy Utilization of Patient Portal for Older Adult Patients

Brittany L. Melton, PhD, PharmD¹, Crystal Burkhardt, PharmD, MBA, BCPS²
¹University of Kansas School of Pharmacy, Lawrence, KS; ²University of Kansas Medical Center, Kansas City, KS

Introduction
Meaningful Use (MU) is part of the effort to increase utilization of healthcare technology, and requires interfaces, such as patient portals, which allow patients to view health records and interact with providers online. In 2014, over 90% of hospitals reported having the capacity for patients to view their health record online, but only 22% of patients reported viewing medical tests online.¹,² Older patients are less likely to enroll with a patient portal and even fewer actively use one.³ Requirements for patient portal use can be met if patients designate a proxy who regularly uses the portal on their behalf. The objective of this study was to evaluate how proxies for older adult patients use a patient portal and how that use impacted their relationship with the patient.

Methodology
The study was conducted at a large teaching medical center which implemented a patient portal in 2011. A list of older adult patients who had enrolled in the patient portal was evaluated for those patients who had designated a proxy for possible inclusion in the study. Patients and their proxies were contacted by phone and asked if they would like to participate in a survey, and verbal consent was obtained. Basic demographic information and relationship with the patient was obtained. The proxy was then asked about how they became involved in the management of the patient’s healthcare information, how often they access the patient portal and what patient information/portal features they use regularly, and which they use least. Participants were also asked how becoming involved in the patient’s healthcare affected their personal relationship. The interviews were transcribed and analyzed for emerging themes. This study was approved by the Institutional Review Board.

Results
Approximately 150,000 patients over 65 years old were seen at the study site between 2011 and 2015. Almost 20,000 older adult patients had accessed the patient portal, and only 39 had designated a proxy in that timeframe. Patients contacted did not know the portal existed and had never used it, therefore they were excluded from the study. Most proxies were female (80%), and Caucasian (80%), with an average age of 60 years. Proxies stated they use the patient portal because their older adult patient was not computer literate or had other limitations. Proxies reported most frequently using laboratory test review, communication with providers, and prescription refill requests. They felt portal use reduced the number of questions they had to ask provider and made healthcare information easier to understand. Proxies reported accessing the portal when they received an e-mail indicating new information was available. Generally proxies felt patient portal use improved their relationship with the patient.

Conclusion
Proxies believe having access to the personal health information for an older adult patient was beneficial for providing care and improved their relationship with the patient. Older adult patients often have cognitive or physical limitations which make patient portal use difficult or impossible, but proxies can serve as a link between the providers and the patient through patient portal access.

References
A Framework for Robust Information Extraction from Free Text Documents

Frank Meng PhD\textsuperscript{1,2}, Craig Morioka PhD\textsuperscript{1,3}, James Sayre PhD\textsuperscript{1,4}
\textsuperscript{1}Medical Imaging Informatics, Dept. of Radiological Sciences, UCLA, Los Angeles, CA
\textsuperscript{2}MAVERIC, VA Boston Healthcare System, Boston, MA
\textsuperscript{3}Dept. of Radiology, VA Greater Los Angeles Healthcare System, Los Angeles, CA
\textsuperscript{4}Dept. of Biostatistics, UCLA, Los Angeles, CA

Clinical NLP has a rich history with many examples of systems that have performed well within laboratory settings [1]. However, most of these systems have not successfully transitioned into operational workflows due to several factors: they are designed mainly for use by informatics experts, they must be re-trained for new extraction problems, and they require relatively large amounts of expert-annotated data for training [2]. In this work, we describe an information extraction (IE) system that attempts to provide an end-to-end solution to address these issues centered around the following goals: 1) generate manually annotated data as a by-product of existing operational tasks using user-centered tools that streamline clinical workflows in order to increase the likelihood of adoption; 2) minimize the amount of annotated data needed to train IE systems to perform at acceptable levels; 3) maximize knowledge re-use to minimize re-training for new extraction tasks; and 4) perform periodic and regular validation on the extracted data to determine the goodness of the data as well as to ensure the continued performance of the system as time progresses.

The overall system architecture is depicted in the figure below and consists of three main elements: the user interface facilitates chart reviews for abstracting data elements from patient records as well as validating automatically extracted data; the IE system automatically extracts data elements from clinical documents using extraction patterns generated from annotated training examples; and the validation and quality assurance process consisting of evaluating extracted data as well as the deployment of alerts to detect performance deviations from expected behavior. The pattern-based IE system was implemented and validated for extracting data from radiology reports of lung cancer patients and was shown to perform better than baseline methods [3]. Because of its reliance on lexical patterns, preliminary results indicate that reasonable extraction performance can be achieved with smaller numbers of training examples. In addition, patterns can be abstracted to represent higher level concepts that potentially make them immediately reusable for other extraction tasks. We designed the user interface as a form-based web application that enables users to efficiently pull specific data elements from patient documents into structured forms without the need for keyboard data entry. The user interface will also enable experts to quickly validate randomly sampled data that was automatically extracted by the IE system. The quality assurance process can be tailored to the amount of resources available as well as the specific goodness level required of the data.

Automated Dynamic Problem and Allergy Lists for Efficient Electronic Health Record Management

Stéphane M. Meystre, MD, PhD\textsuperscript{1,3}, Jianyin Shao, PhD, MS\textsuperscript{1}, Greg Jones, PhD\textsuperscript{2,3}

\textsuperscript{1}Department of Biomedical Informatics, \textsuperscript{2}Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah, \textsuperscript{3}Clinacuity, Inc., Salt Lake City, Utah.

Abstract: To automatically improve the quality of medical problem and allergy lists, and enable more efficient clinical decision support to reduce medication prescription errors, we developed a new prototype to automatically extract structured and coded medical problems and allergies from clinical narrative text in the electronic health record. When evaluated with a focus in cancer outpatient care and a manually annotated corpus of 770 clinical notes, the prototype identified approximately 95% of medical problems, and 90% of allergens.

Introduction: Medical errors are recognized as the cause of numerous deaths, and even if some are difficult to avoid, many are preventable. Computerized physician order-entry systems with decision support have been proposed to reduce these errors, but these systems rely on structured and coded information in the electronic health record (EHR). Unfortunately, a substantial proportion of the information available in the EHR is only mentioned in narrative clinical documents. Electronic lists of problems and allergies are available in most EHRs, but they require manual management by their users, to add new problems, modify existing ones, and the removal of the ones that are irrelevant. Consequently, these electronic lists are often incomplete, inaccurate, and out of date. Our hypothesis is that a system based on natural language processing (NLP) using stepwise hybrid methods can continually extract relevant information from the EHR with high accuracy, to then improve the completeness, correctness, and timeliness of electronic lists of medical problems and allergies. This new system will help ensure the majority of medical problems and allergies of a patient are known by their healthcare providers, and are available for decision support and quality improvement.

Methods: To establish the feasibility of a high accuracy stepwise hybrid NLP system to automatically extract medical problems and allergies from clinical narrative text, we 1) created a reference standard for training and testing, 2) developed the prototype to automatically extract medical problems and allergies, and 3) tested the prototype with the aforementioned reference standard. A web-based problem and allergy lists mock-up resembling the commercial EHR was also developed to prepare the EHR integration. A cohort of 154 patients with breast, lung, or gastrointestinal cancer was randomly selected, and 5 most recent clinical notes of select types (chosen by domain experts) were extracted and de-identified. A new text annotation tool was installed (Brat(1)), tested and then used by 9 experts. Their agreement when annotating was good (F\textsubscript{1}-measure 0.70, or even 0.90 if not considering detailed modifiers), allowing for a good quality reference standard.

The application prototype was developed within the Apache UIMA framework, with text pre-processing analysis engines adapted from Textractor(2) or OpenNLP with trained models from cTAKES. The high-sensitivity component includes a dictionary lookup module and CRF-based named entity recognition.(3) Candidate medical problem and allergies were then filtered by rule-based and SVM-based filters in the false positive filtering components.

Results: For the final evaluation, the prototype was trained and tested using a 10-fold cross-validation approach, and performance reached an average recall of 0.922, precision of 0.960, and F1-measure of 0.940. High-precision filtering in the stepwise hybrid approach allowed eliminating 69\% of the false positive problems identified, increasing precision from 0.673 to 0.921. Precision of allergens identification was already at 1.0, and filtering therefore had no impact (Table 1).

<table>
<thead>
<tr>
<th>Table 1: Prototype evaluation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified problems</td>
</tr>
<tr>
<td>Identified problems</td>
</tr>
<tr>
<td>Filtered problems</td>
</tr>
<tr>
<td>Identified allergies</td>
</tr>
<tr>
<td>Filtered allergies</td>
</tr>
<tr>
<td>Filtered allergens and problems</td>
</tr>
</tbody>
</table>

F1-measure is the harmonic mean of recall and precision (with equal weight for each).

Acknowledgments: Research supported by the National Cancer Institute (R41CA180190).

References

Influenza Vaccine Cancel-Reordering Reveals a Potential Limitation of Immunization Decision Support during Intermittent Vaccine Shortages

Jeremy Michel, MD, MHS;¹ Levon Utidjian, MD, MBI;¹ Jeritt Thayer;¹ Robert Grundmeier, MD;¹
¹The Children’s Hospital of Philadelphia, Philadelphia, PA;

Introduction

Our institution relies on home-grown CDS to support immunization ordering.¹ However, when primary care sites run out of a vaccine, the CDS continues to recommend an out-of-stock product. This is a significant problem for influenza immunization delivery because we use multiple influenza vaccine formulations. Often, clinicians are alerted by nurses that the order must be switched due to a vaccine shortage (creating a cancel-reorder event). As part of an ongoing effort to improve immunization delivery efficiency, we are investigating these cancel-reorder events, their effect on visit duration, and the feasibility integrating real-time vaccine shortage information into the CDS.

Methods

We extracted data on influenza immunization ordering for one clinical care site within our institution from the EHR for the 2015-16 flu season. A cancel-reorder event was recorded whenever a patient had more than one influenza immunization ordered on a given day and at least one of these orders had a status of ‘canceled’. Differences between timestamps were used to calculate reorder-based delay. We are developing a portal to support documentation of out-of-stock events and have modified the immunization CDS to incorporate this data into order recommendations.

Results

A total of 14,572 influenza immunizations were ordered this flu season for 12,934 patients. Of these orders, 19.4% (2,821) were cancelled. Alternate influenza formulations were reordered for 1,496 patients (19.4%) and 58% of patients with a cancelation had reordered immunizations. Reordering occurred 24 minutes on average after the initial order (range 0-200 minutes). Of note, 95 orders were cancelled and reordered within one minute of the initial order.

Discussion

Cancelling and reordering vaccines to change product formulation occurs frequently. More than half of cancelled influenza orders are reordered with an alternate formulation and out-of-stock status is likely the primary cause. Out-of-stock events occur every flu season for formulations, insurances, and care sites. Updating centrally for each event would not be feasible, as the system must be taken offline to update. A web portal (figure 1a) is under development to allow care sites to document local out-of-stock events and share real-time supply data with the CDS (figure 1b).

Figure 1: (a) Prototype portal for nurses to report out-of-stock events. In this example, the practice is out of HPV4 (Gardasil) and Flu Mist. (b) Clinician-facing CDS alters the recommendations to account for vaccine shortages.

Conclusion

Almost 20% of influenza immunization orders are cancelled and 58% of these are reordered with alternate products, likely due to vaccine shortages. Patients wait an average of 24 minutes for reordering to occur, which could affect visit duration. Because central CDS updates in response to vaccine out-of-stock events are not feasible, a portal is being developed and will soon be tested to determine the impact on cancel-reorder events and visit duration.

References

Expert Interpretations of Prostate Cancer Quality-of-Life Survey Results

Sean P. Mikles, MPH1, Andrea L. Hartzler, PhD2, Ted A. Skolarus, MD3, John L. Gore, MD MS1

1University of Washington, Seattle, WA; 2Group Health Cooperative Research Institute, Seattle, WA; 3University of Michigan, Ann Arbor, MI

Introduction

Collecting patient-reported outcomes (PROs) facilitates the monitoring of prostate cancer patients’ quality of life (QOL) after treatment and is considered an indicator of high quality care. Experts have suggested that graphic visualizations can help in the interpretation of PRO scores.1 However, we lack consensus on how best to represent PRO data to support clinical practice. We aim to identify how experts interpret scores from a standard prostate cancer QOL instrument to address this gap and aid in the design of prostate cancer PRO-based visualizations.

Methods

We sent two rounds of electronic questionnaires using a Delphi method to 9 prostate cancer clinician and research experts involved in an international Movember prostate cancer initiative to understand how they interpret Expanded Prostate Cancer Index Composite (EPIC) scores for three prostate cancer domains: bowel, sexual, and urinary function. Participants were asked to rank the importance of the following concepts in clinical practice: whether a patient’s EPIC score is below a threshold value, below a population average, below a baseline score, whether the patient has improved since treatment, and whether the patient is bothered by their QOL. We presented 6 prototype line charts showing an individual patient’s EPIC scores and average scores from a research cohort over time. Each prototype aimed to aid interpretation by showing one of the following types of visual features: 1) confidence interval (CI) around average cohort scores, 2) CI around individual patient scores, and 3) lines defining severity of illness (examples in Figure 1). We asked whether these visual features were useful for clinical decision-making.

![Figure 1. Prototype line charts of prostate cancer urinary incontinence quality of life scores.](image)

Results

After the second round of surveys, the top three averaged concept rankings in each domain were (from most to least important): whether a patient is bothered by their QOL, scores being below baseline, and whether a patient has improved. When considering the prototypes, all 9 experts thought that charts defining illness severity were useful, 9 thought that CIs around cohort scores were useful, and 6 thought that CIs around individual scores were useful.

Conclusion

Participants thought that charts visually defining illness severity are useful for clinical practice, but that threshold scores were not of prime importance. On average, participants thought that it was most important to consider whether patients are bothered by their QOL. Further research is needed to relate prostate cancer QOL scores to clinical decision-making processes to inform future visualizations. This study is limited by its small sample size.

References

The Importance of Well-Defined Mapping Heuristics for Mapping Quality

Catherine Hoang, BSN, MS¹, Holly Miller, BSN, MS¹, Nathan Stocks, MS², Tim Williams, BS³, Paloma Hawry³, Mary Ferramosca, RPh²
¹U.S. Department of Veterans Affairs; ²Deloitte Consulting, LLP; ³Deloitte and Touche, LLP

Introduction

The Department of Veterans Affairs (VA) recognizes the need to modernize legacy systems and to work toward semantic interoperability internally, with the Department of Defense (DoD), and external providers. Due to the number of patients that move between VA and DoD health care systems during their service, information sharing is critical to maintain current and accessible health information, patient safety, continuity of care, and the high quality of care.¹

Section 713 of the National Defense Authorization Act for Fiscal Year 2014 (NDAA) indicates that while VA and DoD share significant amounts of read-only data, much of the data is not standardized, limiting interoperability and clinical decision support.² Semantic interoperability can be achieved through the mapping of prioritized clinical domain data to that of standardized clinical reference terminologies.

Discussion

Mapping to standardized clinical reference terminologies allows semantic interoperability between Departments to be realized without overhaul of the entire system. The most important steps for the initiation of any mapping project are to understand the business and use cases and to establish relevant mapping heuristics.³ Clearly defining the rules that will be applied to not only the project, but any future maintenance of the maps is essential to clarify ambiguity, allow for reproducibility, as well as to provide transparency to the methodology used. It is important to define and test these rules against the actual data set that will be mapped. Our poster will highlight examples of weak and strong heuristics, show how to document these for users to understand the maps, and discuss the implications for map quality and use. Ambiguous and untested rules applied while mapping result in higher disagreement rates between the mapper and validator during the quality review process, and increase the level of effort within an already manual and labor intensive process.

The VA Knowledge Based Systems (KBS) Terminology Standards team has performed validation of mappings within 12 domains and 5 subdomains totaling 121,194 individual terms to support VA’s interoperability goals. Reductions in the disagreement rates were found in each domain as the mapping heuristics were refined. The poster will describe the workflow and map quality validation methods used.

Conclusion

Clearly defined mapping heuristics at the start of a project greatly reduces the ambiguity encountered within the mapping process and amount of effort that may be required to refine or maintain maps.

References


Figure 1: Disagreement Rate
Discovering Community Resources to Support Emerging Models of Care
Mari Millery, PhD & Rita Kukafka, DrPH, MA, FACMI
Columbia University Department of Biomedical Informatics, New York, NY

Background: Emerging healthcare models embrace population outcomes, social determinants, and value-based care. They recognize the role of community resources and supportive services in keeping populations healthy. Engagement of community partners requires accurate information about the complex and dynamic ecosystem of health-related resources. The traditional methods of community asset mapping in public health have not leveraged informatics tools. The goal of this study was to mine multiple available data sources to produce a maximally comprehensive and exhaustive inventory of community health resources in one target community, and to document the steps and data sources required in such effort. The study provides groundwork for partially automating community resource identification in the future.

Methods: The target community of Washington Heights-Inwood (WAHI) in New York City is a densely populated urban area of approximately 280,000 residents; 71% Hispanic and 14% African American. This study is part of a larger GetHealthyHeights.org (GHH) project to implement a community-engaged health informatics (CEHI) platform in WAHI. The resource inventory will be used to populate a publically available resource directory interface on GHH. Use cases to be supported include client referrals, resource access for community members, and a listing of potential collaboration partners. Seven data sources were initially identified through key informants and web searches. Two were excluded due to poor geographic match and outdated information. The remaining five data sources are listed in Table 1, and include online directories, electronic files, and paper-based directories. Table 1 shows resource sectors targeted, geographic coverage, updatability of each data source, and numbers of resource entities found. Key variables were extracted from each source. Non-health entities and duplicates within one source were excluded. Matching of unique entities across sources was conducted manually, using sorting and searching functions. Summary statistics were calculated to examine overlap across sources, unique contributions of sources, and coverage of different resource sectors.

Results: Table 1 shows the numbers of resource entities each data source yielded, after excluding non-health related entities and duplicates within one source (entities kept column). The IRS data was most significantly reduced in this step due to a large number of non-health entities. Before matching for unique entities across sources, the data file had 472 items. After matching, a total of 296 unique entities were found. Among the 296 entities, 58% were found in only one source (single source), 30% in two sources, 7% in three sources, 4% in four sources, and 1% in all five sources. Table 1 shows the number of single source entities per source. All sources yielded single source entities, ranging from 6 to 59. The NYC directory contributed the most single source entities, explained by the fact that it lists many public governmental resources not covered by the other sources. Based on information readily apparent in the sources, it was not possible to code the sector of 38% of entities. The available information suggests robust coverage of the non-profit sector by IRS and HITE, and the public sector by NYC and GHH (Table 1, 3 rightmost columns). The sources included did not provide coverage of the private and informal resource sectors.

Conclusions: The results suggest that it is necessary to utilize a variety of data sources to achieve maximal coverage of resource sectors and entities. The IRS data is available nationally and will provide a source of non-profit entities for any U.S. community. Future steps involve taxonomy-building and semi-automated coding and entity matching.

Table 1. Data Sources and Key Data on Entities Identified

<table>
<thead>
<tr>
<th>Data source</th>
<th>Sectors targeted</th>
<th>Geographic coverage</th>
<th>Date and updatability</th>
<th>No. entities listed</th>
<th>No. entities kept</th>
<th>No. Single source</th>
<th>% Non-profit sector</th>
<th>% Public sector</th>
<th>% Unknown sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRS (990-form)</td>
<td>Non-profit, includes non-health</td>
<td>4 zip codes from state file</td>
<td>Active entities as of Jan 2016, updated monthly</td>
<td>449</td>
<td>67</td>
<td>27</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>New York City Department of City Planning directory of facilities (NYC)</td>
<td>Multiple, including public</td>
<td>Community District 12 (4 zip codes)</td>
<td>2015 directory, updated annually</td>
<td>176</td>
<td>164</td>
<td>59</td>
<td>29.9</td>
<td>35.5</td>
<td>35.0</td>
</tr>
<tr>
<td>Columbia Community Partnership for Health listing (CCPH)</td>
<td>Community-based organizations</td>
<td>Focus on 4 zip codes</td>
<td>2015 directory, one time</td>
<td>34</td>
<td>34</td>
<td>6</td>
<td>71.9</td>
<td>3.1</td>
<td>25.0</td>
</tr>
<tr>
<td>HITE online directory of health and social services in New York City (HITE)</td>
<td>Multiple</td>
<td>Searched for 4 zip codes</td>
<td>Listed entities as of Feb 2016, ongoing updates</td>
<td>100</td>
<td>89</td>
<td>40</td>
<td>47.9</td>
<td>9.9</td>
<td>42.3</td>
</tr>
<tr>
<td>Resources listed on GetHealthyHeights.org (GHH)</td>
<td>Multiple</td>
<td>Primary focus on 4 zip codes</td>
<td>Listed entities as of Feb 2016, ongoing updates</td>
<td>147</td>
<td>118</td>
<td>41</td>
<td>25.0</td>
<td>41.7</td>
<td>33.3</td>
</tr>
</tbody>
</table>
Applying Machine Learning Methods to Predict Activities of Daily Living for Cancer Patients

Hua Min, Talha Oz, Sava Vukomanovic, Hedyeh Mobahi, Katherine Irvin, Ilirjeta Krasniqi, Janusz Wojtusiak
Department of Health Administration and Policy, George Mason University, Fairfax, VA

Introduction
The ability to predict Activities of Daily Living (ADLs), their changes, and relationships to comorbid conditions are of great importance for clinicians, patients, and caregivers. In recent years, researchers have applied machine learning techniques to create models capable of predicting ADLs for patients. However, unexplored factors remain including the effect of multiple comorbidities and their combinations on ADLs. In this study we applied machine learning methods including AQ21 and WEKA to analyze the SEER-MHOS data. SEER-MHOS links two large population based sources of data. SEER data comes from cancer registries, that collect clinical information, demographics, and cause of death. MHOS provides information about patient physical and mental status obtained in clinical surveys.

Methods
SEER-MHOS data have been used to extract comorbidities and ADLs (self-reported), as well as cancer characteristics (SEER registry). The total number of patients in SEER-MHOS data is 1,849,311 we then extracted only cancer patients 102,269. Furthermore, patients with multiple cancers were excluded resulting in 83,255 patients left. Then we extracted patients who completed both surveys 2-year before and 1-year after the cancer diagnosis. If a patient had multiple surveys, one survey before and one survey after closest to the cancer diagnosis date were used. The final data contain 2,239 cancer patients. This study included six ADLs (walking, dressing, bathing, moving in/out chair, toileting, and eating) measured before and after cancer diagnosis, 12 comorbidities (e.g., Hypertension, Diabetes, etc.) before cancer diagnosis, six cancer characteristics (grade, staging, tumor size, histology, tumor extension, and behavior), radiation and surgery treatment indicators, and three demographics (age, race, and marital status). Models have been developed using AQ21 to predict ADL functionalities after cancer diagnosis. AQ21 is a supervised machine learning system for creating attributional rules for data and background knowledge. The software has been used to discover rules that indicate strong relationships in data, but grouping patient characteristics that are indicative of deficiencies in ADLs. The data have also been analyzed by using WEKA software to create baseline accuracy from methods other than AQ21. Specifically, we tested Logistic Regression, Decision Tree, Random Forrest, Naïve Bayesian, all with 10-fold cross validation.

Results
There are 23 rules generated by the AQ21 for the six ADLs. The following shows one sample rule.
Rule1: [Bathing_disability] ==>[Histology = 2,1: 45, 80, 36%, 45, 80, 36%]
[Staging = 7: 40, 72, 35%, 21, 25, 45%]
[Primary site = Lung and bronchus, Colon, Large intestine: 53, 109, 32%, 20, 14, 58%]
:p = 20, n = 14, q = 0.315, cx = 27
First line in the rule indicates that among the patients with histology type of epithelial neoplasms and squamous cell neoplasms, 45 of them have bathing disability while 80 patients have not. The consistency of the prediction is 36%.

Discussion
The machine learning methods can be used to identify predictors for ADLs for cancer patients. When applying such methods one needs to tradeoff accuracy of the learned models for interpretability and transparency of results. In general, rules are known to be the most transparent representation, directly corresponding to statements in natural language. Since AQ21 cannot calculate AUC for rules it learns, we are currently working on other ways of comparing the results of AQ21 and WEKA.

Acknowledgments
The study was supported by the Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee.
Visualizing the Effects of Cancers on Relationships Between Comorbidities and Activities of Daily Living

Hua Min, Talha Oz, Sava Vukomanovic, Hedyeh Mobahi, Katherine Irvin, Ilirjeta Krasniqi, Janusz Wojtusiak
Department of Health Administration and Policy, George Mason University, Fairfax, VA

Introduction
Visualization methods are useful in helping researchers discover and explain relationships in data. They can offer new ways to view and present data insights to users faster than traditional analysis methods. In the presented work we use visualization techniques to explore Activities of Daily Living that are important indicators of quality of life. The ability to predict ADLs, their changes, and relationships to comorbid conditions is of great importance for clinicians, patients, and caregivers. In recent years, researchers have applied machine learning techniques to create models capable of predicting ADLs for patients. However, unexplored factors remain including the effect of cancers on comorbidities and ADL relationships. Our hypothesis is that for some patients, the presence of specific cancers may strengthen the impact comorbidities have on ADLs, and for others it may weaken the relationship.

Methods
SEER-MHOS data have been used to extract comorbidities and ADL measurements (self-reported), as well as cancer characteristics (SEER registry). The data consisted of 102,269 cancer and 1,747,042 non-cancer patients. This study included 12 self-reported comorbidities (e.g., Diabetes, Hypertension, etc.) and six ADLs (e.g., walking, bathing, dressing, etc.) from 10 cancers (e.g., Prostate, Breast, Colorectal, etc.). The data have been loaded to PostgreSQL database and pre-processed to extract needed variables. Statistical calculations have been done in Python using SciPy library, and visualization using Gephi. First relationships between comorbidities and ADLs were analyzed and the resulting comorbidity and ADL relationships were paired (comorbidity, ADL). Strength of relationships between pairs was measured as likelihood ratio \( LR_{comorbility,ADL} = \frac{p(comorbility|ADL)}{p(comorbility|no ADL)} \), which was then compared between non-cancer and cancer patients by cancer type. The rationale between the two networks is that figure 1 focuses on direct relationships between ADLs and each comorbidity, while figure 2 examines effects of cancer types on ADLs given a comorbidity. Fisher’s exact test has been used to calculate statistical significance of the effect of the cancers on the paired relationships, and to graph those that were statistically significant (p<0.01). The strength of the connections have been calculated as \( |LR_{comorbility,ADL} - LR_{comorbility,ADL,nc}| \), the difference between likelihood ratios for cancer and non-cancer patients, and edges with strength > 0.51 remain.

Results
Figure 1 shows the relationship between comorbidities (outside nodes) and ADLs (inside nodes). The thicker the connection the more often patients with a comorbidity reported an impact on that ADL. Congestive Heart Failure (CHF), Stroke, and Gastrointestinal disease (GI) comorbidities impact the most ADLs while Hypertension the least, with CHF to walking have the strongest weighted relationship (3.08) and Hypertension to eating the weakest (1.05). Figure 2 pairs the relationships from Figure 1 and presents the effect that cancers have. The smaller circles are cancers and are connected to 68 comorbidities and ADL pairs. Pancreatic cancer not only impacts the most number of pairs but the strength of the relationships are more significant. For example, when you compare Pancreatic and Stomach cancers impacts on the paired node GI disease and dressing ADL, it was found the weighted strength of Pancreatic is 1.37, whereas Stomach is 0.61.

Discussion
The visualization of the SEER-MHOS data discovers not only the complicated relationships among comorbidities, ADLs, and cancers but also the weights of those connections. Findings from the figures are clinically meaningful. For example, Pancreatic cancer which is one of the most aggressive cancers has the strongest impact on ADLs. Our next steps for this study are discovering relationships among cancer characteristics, comorbidities, and ADLs.

Acknowledgments
The study was supported by the Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee.
Development of a Diagnostic Assistant for Multiple Sclerosis Using Natural Language Processing of the Electronic Medical Record

Lindsey R. Mitrani, BA\(^1\), Gabriel G. Lu\(^1\), Dominick J. Fulgieri\(^1\), Herbert S. Chase, MD, MA\(^1\)

\(^1\)Columbia University Medical Center, New York, NY

Introduction
Missed or delayed diagnoses are a major cause of medical error. Diagnosis could be facilitated by “diagnostic assistants,” algorithms which mine the electronic health records (EHR) for clues of one or more conditions. For diseases characterized by abnormal lab values, such as chronic kidney disease, implementing a diagnostic assistant that surveys the patients’ structured lab data is computationally simple (1). However, for diseases for which there are no routine lab tests and are instead characterized by signs and symptoms such as neurological illnesses, the diagnostic assistant would have to mine unstructured data using natural language processing (NLP).

Objective
The goal of this study was to build a diagnostic assistant for Multiple Sclerosis (MS), using NLP to parse clinical notes, which could identify patients with MS. We chose MS given the importance of prompt intervention in mitigating progression of the disease and the known delay in diagnosis (2, 3).

Materials and Methods
Clinical notes of patients from the adult outpatient clinic were extracted from the Columbia University Medical Center clinical data warehouse. The notes were parsed using MedLEE which maps signs and symptoms to UMLS terms (4). The presence of an ICD9 code for MS was the gold standard for identifying MS patients. Training set data consisted of parsed clinical notes of MS patients who had the illness for at least two years and of controls. We used this data set to identify UMLS terms associated with MS and to build the classification model. Test set data consisted of parsed notes of randomly selected patients from the same adult outpatient clinic, some of whom had a MS. Classification of the training set was performed using the Naive Bayes classifier using 10-fold cross validation and the Weka workbench. Classification of the test set used the training set to generate a model.

Results
Approximately 1000 UMLS terms from the notes of MS patients of the training set occurred more frequently than in controls. Synonymous terms were manually clustered into 67 separate buckets reflecting different characteristic signs and symptoms of MS, such as “paresthesia,” and used as features in classification. If a patient had any one of the bucket terms in a single note they were scored a 1 for that bucket; if not, a 0. Classification using the training set yielded excellent accuracy (Table). When the model was applied to the larger test set, the accuracy of classification was similar. The post-test probability of MS in the test set increased five fold from 1.2% to 6.0%.

Discussion and Conclusion
Our results demonstrate that a classification model that accurately identifies patients with MS can be developed using signs and symptoms of MS extracted from clinical notes using NLP. Although the post-test probability increased 5-fold, the low prevalence of MS and the 87% specificity of classification resulted in false positives. Some of these patients, however, might have MS that was not recognized by their providers. A manual review is underway to explore this possibility. Errors of ICD coding for MS likely reduced classification accuracy, as well.

References
Predictive Modeling for Appointment No-show in Community Health Centers

Iman Mohammadi, MS1, Ayten Turkcan, PhD2, Tammy Toscos, PhD1,3, Huanmei Wu, PhD1, Brad N. Doebbeling, MD, MSc4

1Indiana University, Indianapolis, IN; 2Northeastern University, Boston, MA; 3Parkview Research Center, Fort Wayne, IN; 4Arizona State University, Phoenix, AZ

Problem. One key measure for improving access to care is reducing the number of “no-shows.” An appointment is considered a no-show when the patient misses the appointment without cancelling. No-show rates from 10% to 50% have been reported in different healthcare settings1,2. Purpose. In our current multiyear and multisite project, we are using simulation modeling to test and find the optimal scheduling processes, staffing and policies for improving access to care2,3. Thus, we estimate patients’ no-show probabilities using statistical methods such as logistic regression. Here, we present a no-show prediction model and report predictors of no-shows by analyzing Electronic Medical Record (EMR) and scheduling data. Methods. We evaluated different factors which we hypothesized may be related to appointment no-show. We collected all appointments and patient encounters during 2014 from an urban Community Health Center (CHC) in Indiana. Data fields included appointment characteristics (date of visit, date the patient had contacted the clinic to arrange an appointment. To test the effect of patient no-show behavior on appointment adherence, we created another variable “prior no-show rate” for each patient. Prior no show rate is the number of no-shows for a given patient prior to the last appointment, divided by the patient’s total number of appointments prior to the last appointment. Appointment adherence (“no-show” or “arrived”), a binary variable, was the outcome variable in this project. We developed the no-show prediction model using logistic regression analysis in SAS 9.4 and also tested variables individually against the appointment compliance. Results. The multivariate logistic regression model considered all variables in the dataset plus lead-time and patient prior no-show rate. Predictors of no-show in our final model included provider specialty, insurance type, age, cellphone availability, lead-time, and patient prior no-show rate. The overall no-show rate was 12%, and the c-statistic for the model was 0.76. Approximations scheduled with pediatricians and gynecologists had higher no-show rates, compared to those with nurse practitioners and internal medicine physicians (p <0.0001). Patients scheduled with pediatricians (OR=4.3, 95%CI=[1.6,6.6]) or with OB-GYN (OR=3.2, 95%CI=[2.2,8.6]) were more likely to no-show than those scheduled with internists. Self-pay and Medicaid patients had higher no-show rates than private insurance or Medicare patients did (p <0.0001). Patients who did not have cell phones had no-show rate of 45% while they had arrival rate of 35% (p=0.005). Patients without cellphone were 1.5 times more likely to no-show than patients who have a cellphone (95%CI=[1.2,1.9]). The univariate tests show that the average lead-time of 22.8 days for “no-show” appointments is larger than the average lead-time of 17 days for “arrived” appointments (p <0.0001). Compared to same-day appointments, appointments made less than 2 weeks in advance, between 2 weeks and 1 month in advance, and more than 1 month in advance are respectively 5.1 (95%CI=[3.3,8]), 4.6 (95%CI=[2.8,7.5]) and 7.5 (95%CI=[4.7,12]) times more likely to become no-show. The mean of prior no-show rates for “no-show” appointments at 11% is larger than the mean of prior no-show rates for “arrived” appointments at 4% (p <0.0001). That means, chronic no-showers are more likely to no-show (OR=5.6, 95%CI=[3.2,6.8]). Other variables, appointments time (AM vs. PM), gender, weekday, ethnicity, race and marital status, did not impact appointment adherence. Discussion. There are three key findings. First, more same-day appointments (greater open access scheduling) can reduce clinics’ no-show rates. It has been suggested that patients might miss their appointments made far in advance because of forgetting the appointment, getting better, or having other priorities1. Second, patients who frequently not shown for appointments in the past are more likely to no-show again. This group might be limited to same-day appointments. Third, higher no-show rates among patients without a cellphone may be a result of reduced appointment reminders. Alternative means for scheduling and appointment reminders may reduce no-show rates and improve access to care. Conclusion. EMR and scheduling data is useful for no-show prediction models, which can inform decisions about system redesign. Patient and provider characteristics and visit features help predict appointment adherence.

References

Semantic Relatedness and Similarity between Biomedical Concepts

Sungrim Moon, Ph.D., Trevor Cohen, Ph.D., Hua Xu, Ph.D.
School of Biomedical Informatics, University of Texas Health Science Center Houston, Houston, TX, 77030, USA

Introduction: Automatic differentiation of semantic relatedness and similarity among biomedical concepts has been explored by using path-based, Information Content (IC), lexical matching, and semantic vector (or distributional) methods in the biomedical and clinical domains. Nevertheless, automated discovery of grouping semantic relatedness and similarity concepts is still challenging, due to dependence on domain knowledge and limited word-coverage. We used the distributional semantics approach with relevant Named Entity Recognition (NER) to automatically distinguish concepts of relatedness and similarity from a large unlabeled corpus.

Method: This study focused on semantic vector approaches using three different corpora. We used one clinical corpus from the Multiparameter Intelligen Monitoring in Intensive Care (MIMIC) II data (403,871 notes) as well as another set of diverse clinical narratives from the University of Texas Health Science Center Houston Clinical Data Warehouse (911,955 notes). We also used the titles and abstracts (n=1,249,980) from the 2013 edition of MEDLINE. We processed all data to separate sentences and tokens, and extracted UMLS concepts (CUIs) using Clinical Language Annotation Modeling and Processing (CLAMP)\(^1\). CLAMP encoded sentences as sets of problems, treatments, and laboratory test concepts. We trained a distributional semantics model (variants of Random Indexing using the Semantic Vectors software package\(^2\)) with a range of parameter settings for each corpus – e.g. the Random Direction model is a sliding window-based model that takes into account whether a context term occurs before or after a term of interest. We evaluated correlation with human judgments of similarity and relatedness on pairwise terms\(^3\) using Spearman’s rank correlation. We compared our results with those obtained by path-based, IC, lexical matching, and word-based semantic vector methods, which were re-calculated by McInnes and Pedersen (2015)\(^3\).

Results: Table 1 shows the results of the three distributional semantics models compared to other different approaches. The Random Direction model with semantic NER encoding using CLAMP (MEDLINE 2013 corpus) yielded best correlations for similarity (0.5828) and relatedness (0.5336). Additionally, we tested semantic NER encoding using MetaMap. Though MetaMap produced lower correlation (0.5687 for similarity and 0.4874 for relatedness) than CLAMP, results obtained with MetaMap outperformed all other models tested. This strongly suggests that concept extraction and normalization can improve the correlation between semantic vector models and human judgment of similarity and relatedness, as compared with term-based models.

Table 1. Correlation comparisons. (Best results in bold.)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Similarity</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>path</td>
<td>0.5335</td>
<td>0.3062</td>
</tr>
<tr>
<td>wup</td>
<td>0.5079</td>
<td>0.2633</td>
</tr>
<tr>
<td>lch</td>
<td>0.5335</td>
<td>0.3062</td>
</tr>
<tr>
<td>res</td>
<td>0.4905</td>
<td>0.2800</td>
</tr>
<tr>
<td>jcn</td>
<td>0.5267</td>
<td>0.3555</td>
</tr>
<tr>
<td>lin</td>
<td>0.5137</td>
<td>0.3114</td>
</tr>
<tr>
<td>t-res</td>
<td>0.4910</td>
<td>0.2797</td>
</tr>
<tr>
<td>t-jcn</td>
<td>0.5154</td>
<td>0.3081</td>
</tr>
<tr>
<td>t-lin</td>
<td>0.5052</td>
<td>0.3023</td>
</tr>
<tr>
<td>leks</td>
<td>0.5246</td>
<td>0.4379</td>
</tr>
<tr>
<td>vector</td>
<td>0.5289</td>
<td>0.4048</td>
</tr>
<tr>
<td>semantic vector with NER concept encoding (MIMIC)</td>
<td>0.4386</td>
<td>0.3767</td>
</tr>
<tr>
<td>semantic vector with NER concept encoding (UTHealth)</td>
<td>0.4152</td>
<td>0.3724</td>
</tr>
<tr>
<td>semantic vector with NER concept encoding (MEDLINE)</td>
<td><strong>0.5828</strong></td>
<td><strong>0.5336</strong></td>
</tr>
</tbody>
</table>

Conclusion: With NER encoding using CLAMP, distributional semantics models outperformed the state-of-the-art measurements. It suggested that relevant semantic NER encoding is essential to achieve high-correlations to human judgment. Our method could automatically categorize semantic meanings in a practical way.

Acknowledgement: This study was supported by grant from the NLM R01LM010681.

References

Resolving Hierarchical Ambiguity in Indexing Recommendations

James G. Mork, MSc, Dina Demner-Fushman, MD, PhD,
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD

Introduction
Finding the main points of a given text and mapping them to MeSH® facilitates MEDLINE® indexing, assignment of key terms by authors, cataloguing, etc. One of the problems in assigning terms using the MeSH hierarchies is to decide on the level of specificity that is required for each document. The decision is based on the context of the document, as well as the nature of the indexing task, whether a more general term, a more specific term, or both levels of terms should be assigned. For the task of assisting the MEDLINE indexer, the decision could be based on the indexing rule of assigning the most specific MeSH term appropriate to the article, except in cases where the indexer’s expert judgement is to apply an allowed exception. In this work, we rely on a corpus-based approach to learn how to select an appropriate specificity of the term when our tool, The NLM Medical Text Indexer (MTI) suggests both a general and a specific MeSH term from the same MeSH tree.

Methods
We used 763,227 citations that were indexed in 2015 and also had MTI recommendations (henceforth referred to as Corpus). MTI recommended 6,465,133 MeSH terms and the human indexers assigned 8,454,900 MeSH terms to the Corpus. We focused on citations for which MTI recommended both a more general MeSH term and a more specific MeSH term from the same MeSH tree – for example, Vaccines (D20.215.894) is more general than Anthrax Vaccines (D20.215.894.135.063). When MTI recommends both of these terms for a given citation, we call it Hierarchical Ambiguity. In the Corpus, there are 985,738 (15.25% of MTI total) Hierarchical Ambiguity pairings from MTI and 726,861 (8.60% of human total) from the human indexing, or almost half the rate of occurrence we see from MTI.

Results
The easy solution would have been to just not recommend the more general term per the rule of indexing the most specific term. The problem is that the exceptions noted earlier still constitute 8.60% of the human indexing. In our Corpus, we would lose 314,869 (31.94%) of the correct general term recommendations by always recommending only the most specific terms. Table 1 shows the distribution of indexer assignments where MTI assigned both terms.

Table 1. Human indexing as judgements for MTI recommended Hierarchical Ambiguity.

<table>
<thead>
<tr>
<th>Term Count</th>
<th>Gen Wrong/Spec Right</th>
<th>Gen &amp; Spec Wrong</th>
<th>Gen Right/Spec Wrong</th>
<th>Gen &amp; Spec Right</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen Wrong/Spec Right</td>
<td>426,242</td>
<td>244,627</td>
<td>186,503</td>
<td>128,366</td>
<td>985,738</td>
</tr>
<tr>
<td>% of Suggestions</td>
<td>43.24%</td>
<td>24.82%</td>
<td>18.92%</td>
<td>13.02%</td>
<td>15.25%</td>
</tr>
</tbody>
</table>

Ignoring MTI recommended pairs in which both the general and specific recommendations were wrong in the Corpus, we identified 6,399 Hierarchical Ambiguity pairs in the remainder of the Corpus where the general term was wrong on average 82.88% of the time accounting for 52.40% (223,361) of the Gen Wrong/Spec Right recommendations in Table 1. A simple rule to always remove the general term recommendation when one of these 6,399 Hierarchical Ambiguity pairs is identified in the MTI results, provided us with an improvement in Precision from 0.6292 to 0.6406 (+1.81%) while only dropping Recall from 0.6463 to 0.6419 (-0.68%) for our Test Collection.

Conclusion
Understanding where the problematic Hierarchical Ambiguities are in the MTI recommendations has provided us with a way of eliminating over 50% of the erroneous general term results with very little loss to Recall.

Acknowledgments
This work was supported by the intramural research program of the NIH, U. S. National Library of Medicine.

References
Apache Spark for the Analysis of High Frequency Neurointensive Care Unit Data: Preliminary Comparison of Scala vs. R.

Laura Moss, PhD, MSc, BSc1, Martin Shaw PhD, BSc2 Ian Piper, PhD, BSc2, Chris Hawthorne, MBChB1, John Kinsella, MD, MB, BS1, Aridha, Philips Healthcare
1Dept. of Anaesthesia, Pain & Critical Care, University of Glasgow, Glasgow, UK; 2Dept. of Clinical Physics & Bioengineering, Institute of Neurological Sciences, NHS Greater Glasgow & Clyde, Glasgow, UK

Introduction

Advances in patient monitoring equipment have led to an increasing amount of complex patient data being stored and collected in Neurointensive Care Units. Additionally, the future use of genomic and proteomic data at the patient’s bedside increases the requirements for technology and tools which can support the bedside analysis of this ‘big data’ to inform clinical practice. In the CHART-ADAPT project we are developing a framework which uses the Apache Spark processing engine to enable physiological models to be applied to high frequency waveform data typically collected in Neurointensive Care Units. Spark is well suited to low latency processing and it is envisaged that it will offer performance advantages required for the analysis of high frequency data in clinically meaningful timescales compared to existing critical care data analysis platforms1.

The CHART-ADAPT Framework

The CHART-ADAPT framework receives HL7 patient data, de-identifies the patient data, then transfers it to cloud computing services for configurable big data analytics, the results from the analyses are re-identified and presented back into the clinical environment. This framework is customized for use by Neurointensive Care Units and a pilot study is being conducted at the Institute of Neurological Sciences, Glasgow. In the pilot study physiological patient data is automatically extracted from Philips ICCA and iXTrend monitoring equipment; sampling rates range from 60Hz to 500Hz. Additionally, waveform data for 6 channels (e.g. intracranial pressure (ICP)) is converted from proprietary file formats into MFER and the entire flow of patient data is automatically transformed into HL7. The data is de-identified and transferred to a cloud based service provision including Spark, Hadoop and Greenplum. A number of Neurointensive Care physiological models will be implemented in Scala; results from continuously running the models will be integrated back into the Philips ICCA system and displayed at the relevant bedside.

Preliminary Comparison of Scala Vs. R.

To evaluate the performance gains of the Spark/Scala implementation on high frequency Neurointensive care data, a preliminary study has been conducted to compare the implementation of two models – the pressure reactivity index (PRx)2 and a hypotension area burden model. 7 different datasets were created from existing patient data, these consisted of: 100,000, 200,000, 300,000, 400,000, 500,000, 600,000 and 750,000 rows of ICP and blood pressure data. Firstly, an algorithm was written in R and Scala to down-sample data from the initial sampling rate of 0.01 seconds to 10 seconds. Secondly, both models were implemented in R and Scala and each experiment was repeated 35 times to assess variance in run time. Performance was then compared using the different sized datasets. For the datasets used, the run time of both the R and Scala algorithms increased approximately linearly with data volume. Scala performed in the range of 50 to 200 times faster than the R code across these experiments for both models. The R run times were seen to increase with both the data volume and the overall duration (for the PRx algorithm) while the Scala run times increased only with data volume (for both models).

Discussion

This preliminary study has indicated that the use of Spark/Scala offers performance enhancements for the Neurointensive care unit data. Future work includes evaluation of the other physiological models. Full evaluation of the CHART-ADAPT framework in-practice is planned.

References


Danielle L. Mowery, PhD1, Kristina Doing-Harris, PhD2, Wendy W. Chapman, PhD1, Chrissy Daniels, MS1, Mike Conway, PhD1
1University of Utah, Salt Lake City, UT; 2Westminster College, Salt Lake City, UT

Introduction

Health care providers are increasingly concerned with quality in health care delivery, particularly patient satisfaction. Press-Ganey surveys are distributed electronically as a means of gauging patient satisfaction. However, rich, qualitative descriptions of their experiences are locked within the response free-text comments. Sentiment terms within these comments can convey the emotional valence of an experience with degrees of strength and types of polarity suggestive of a patient’s satisfaction or dissatisfaction. For example, in “service was terrific”, the term “terrific” conveys a strength: strongly and polarity: positive suggesting patient satisfaction. Quality improvement abstractors could benefit from visualizations of survey responses based on emotional valence to identify areas of patient satisfaction and dissatisfaction. In this study, we aimed to characterize the types of terms, strength, and polarities observed from Press-Ganey patient satisfaction survey responses.

Methods

In this Institutional Review Board-approved, pilot study, we sampled 60,697 free-text, Press-Ganey patient satisfaction survey responses from the University of Utah Healthcare System from the year 2014. For each response, we encoded observed sentiment terms and their strength/polarity from the Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon1. We report the frequency of sentiment terms for each strength/polarity class and present visualizations of observed terms by frequency for strongly positive, neutral, and strongly negative responses2.

Results

We observed variable frequencies of unique sentiment terms: strongly positive (n=518 terms), weakly positive (n=435 terms), neutral (n=268 terms), weakly negative (n=397 terms), and strongly negative (n=616 terms).

Figure 1. Word clouds of terms associated with strongly positive (left), neutral (middle), and strongly negative responses (right). Larger word size represents more frequent terms; smaller word size represents less frequent terms.

Conclusion

The MPQA subjectivity lexicon can provide a “bird’s-eye” view of emotional valence of patient satisfaction and dissatisfaction. We are leveraging these annotations as features within a supervised machine learning approach that predicts the sentiment and reason for a patient’s satisfaction or dissatisfaction with his/her hospital care experience.

References

Characterizing the Fever Effect in Autism Spectrum Disorder

Efrat Mullera, BSc, Alal Era nb,c,d, PhD, Denis Agnielb, PhD,
Isaac S Kohaneb,c, MD, PhD, Eitan Bachmatb, PhD

a Department of Computer Science, Ben Gurion University, Beer Sheva, Israel;
b Department of Biomedical Informatics, Harvard Medical School, Boston, MA;
c Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA;
d Department of Life Sciences, Ben Gurion University, Beer Sheva, Israel.

Background

Several anecdotal reports by parents and caregivers suggest that some children with autism spectrum disorder (ASD) demonstrate marked behavioral improvements during febrile episodes. However, the magnitude, scope, and nature of this so-called Fever Effect remain unknown.

Objectives

Improve our understanding of the Fever Effect in ASD by identifying clinical and behavioral characteristics of individuals with ASD reported to improve during febrile episodes, characterizing the domains of improvement, and identifying potential confounding factors.

Methods

We mined medical histories and behavioral data from 2,253 families with ASD from the Simons Simplex Collection. Continuous variables were clustered based on pairwise correlations and categorical variables based on pairwise associations. The relationship of each cluster with fever response was then tested using non-parametric statistics. Multivariate models were constructed based on forward feature selection, using logistic regression.

Results

16.7% [15.2%, 18.3%] of the children included in the analysis were reported to demonstrate behavioral improvements across multiple ASD domains when febrile. They tend to be on the low end of the spectrum, with lower IQ (p < 1x10^{-6}), worse social and communication skills (p < 1x10^{-6}), and increased repetitive and restricted behaviors (p = 1.1x10^{-5}) as compared to children with ASD that showed no behavioral improvements during febrile episodes (Figure 1). Significant associations were detected between fever response and gastrointestinal (GI) dysfunction (OR = 1.45 [1.21, 1.74]), milk, casein and gluten allergies (OR = 2.2 [1.68, 2.91], 3.1 [1.96, 4.96], 2.28 [1.43, 3.52], respectively), and maternal infection in pregnancy (OR = 1.7 [1.42, 2.03]). Maternal cognitive abilities were identified as a confounder, with mothers scoring higher on the Social Responsiveness Scale (SRS) being more likely to observe a fever response in their affected child (p = 3.98x10^{-3}).

Conclusion

An emerging subtype of fever-sensitive ASD may be characterized by GI abnormalities, allergies, and maternal infection in pregnancy. Fever response spans all ASD domains and is more likely to be reported in children on the lower end of the spectrum. Follow-up prospective studies are needed to minimize parental reporting biases.

Figure 1. Cumulative distribution function (CDF) differences in (A) Verbal IQ, (B) Communication skills, (C) Social interaction, and (D) Restricted and repetitive behavior between children with ASD reported to improve when febrile (orange) and those whose behavior reportedly remains unchanged (blue). (B), (C) and (D) are based on the Autism Diagnostic Interview Revised (ADI-R), where higher scores represent severe symptoms.
Validation of an Automated Process for the Comparison of Nutrition Care with Evidence-Based Nutrition Practice Guidelines (EBNPG)
William J Murphy, MS, Academy of Nutrition and Dietetics, Chicago, IL; Martin M Yadrick, MBI, MS, Computrition, Inc., West Hills, CA; Rosa K Hand, MS, Academy of Nutrition and Dietetics, Chicago, IL

Background: This project sought to evaluate a process designed for the Prevention of Diabetes Guidelines Impact Study (PDGIS) to assess the similarity of practiced nutrition care to EBNPG recommendations. PDGIS convened an expert panel to apply a framework that describes nutrition care as 6 connected steps and translate recommendations into expected care plans (ECPs) using the Nutrition Care Process and Terminology (NCPT) standardized language to define expected documentation for each step. PDGIS will compare ECPs with care documented via the Web-based Academy of Nutrition and Dietetics Health Informatics Infrastructure (ANDHII) with an automated process that includes matching based on the use of standardized terminology and natural language processing (NLP) of unstructured text. Our NLP algorithm uses the Princeton University WordNet lexical database to find synonyms of keywords stemmed with Porter's Algorithm and reports a match when at least half of the keywords from the ECP share a synonym with the related documentation. To evaluate this automated process, we compared its results with those of a manual audit.

Methodology: One investigator who was not involved in the development of the ECPs served as the expert reviewer and examined the ANDHII record of encounters from PDGIS for similarity with ECPs, recording a judgment as to whether each step of an ECP was completed. The intra-class correlation coefficient (ICC) was used to evaluate the similarity between the automated and manual processes, and we hypothesized that the two would have good agreement (ICC > .60). A sample size of 23 classifications of success for a step was necessary to differentiate good agreement from very poor agreement (ICC = .1) with 80% power. We sampled 24 initial visit encounters from PDGIS via a block-randomization procedure that sampled 1-2 encounters from each of the 14 participating dietitians. One ECP (n = 13) was randomly assigned for review with each encounter.

Results: For 124 step classifications, the overall ICC was 0.60 and was significantly different from poor agreement (p < 0.001). Figure 1 presents the rates of successful matching to the ECP for each process, the ICC between the two processes by step, and labels for which steps are matched through the use of standardized terms versus NLP keyword matching.

Conclusion: While the overall ICC was at the threshold of our goal of good agreement, large variations appeared between steps. The step of setting goals for treatment outcomes, which relies heavily on NLP for concepts not standardized in terminology, exhibited virtually no agreement despite similar success rates. Our automated process for matching based on standardized terminology is adequate for analysis of practiced care. However, our NLP algorithm for those steps not yet standardized by the NCPT was not sufficiently accurate for use in PDGIS, and we will need to either improve the algorithm, for example through the use of a medical lexicon such as the Unified Medical Language System’s SPECIALIST Lexicon, or manually review these steps.

References
Electronic Health Record Interventions Improve Patient Centered Medical Home Documentation for Primary Care Practices Seeking National Committee for Quality Assurance Recognition

Sara R. G. Myers, BA¹, Elizabeth R. Silvers, BA¹, Catherine L. Liang, MBA, MPH¹, Lynn A. Volk, MHS¹, Colleen E. Blanchette, BA¹, Amy L. Feeney, MBA¹

¹Partners HealthCare, Wellesley, MA, USA

Introduction: The Patient Centered Medical Home (PCMH) model, originating from a pediatric model of care, seeks to improve care coordination and self-management, particularly for patients with complex, high-risk health conditions. The process for primary care practices seeking National Committee for Quality Assurance (NCQA) PCMH recognition is rigorous, and satisfying chart review documentation requirements was challenging for our 2014 cohort of practices. Based on feedback from these practices, we implemented Electronic Health Record (EHR) interventions and developed new educational materials to support documentation efforts for our 2015 cohort. This study aims to determine whether chart review scores changed from our 2014 cohort to our 2015 cohort.

Methods: We conducted a comparative cohort study to assess changes in chart review scores for primary care practices undergoing PCMH transformation in 2014 and 2015. Because the applicable NCQA standards changed between these two years, we evaluated two factors for which the documentation standards did not change and for which less than 90% of practices passed: documentation of treatment goals and providing information about new medications to patients. In order for a practice to meet these two requirements, at least 75% of charts must have clearly documented treatment goals and at least 80% of charts must have included documentation indicating that information about new medications was given to the patient. We also compared scores of pediatric and adult practices within and between the 2014 and 2015 cohorts.

Results: Table 1. Proportion of adult and pediatric practices meeting PCMH documentation requirements in the 2014 and 2015 cohorts (n=number of practices).

<table>
<thead>
<tr>
<th>Documentation Requirement</th>
<th>Adult</th>
<th>Pediatric</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=28)</td>
<td>(n=17)</td>
<td>(n=5)</td>
</tr>
<tr>
<td>Treatment goals</td>
<td>82%</td>
<td>88%</td>
<td>60%</td>
</tr>
<tr>
<td>Information about new medications given to patient</td>
<td>46%</td>
<td>71%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Scores of total practices passing these two documentation requirements improved between the 2014 and 2015 cohorts. Pediatric practices showed greater improvement than adult practices on both of the documentation factors evaluated.

Conclusion: Improved scores for two PCMH documentation requirements may be attributed to EHR interventions and support. Primary care practices in the Partners network use multiple EHR systems, so interventions were specific to EHR functionalities. These interventions included templates in free-text notes and “Patient Educated” checkboxes that primary care providers could check off to indicate that they provided information about new medications to patients. These interventions were easily incorporated into existing workflow, making documentation requirements less taxing for practices. Pediatric practices improved more than adult practices, perhaps due to an emphasis in pediatrics on harnessing support from family and school resources and developing self-management plans that focus on continuity of care. Practices seeking PCMH NCQA recognition could improve their chances of recognition by implementing EHR interventions that support PCMH documentation efforts and are easily incorporated into existing workflow.
Investigation of clinical process visualization using EMR data in clinics

Kodai Nakajima¹, Satoshi Tamura, PhD¹, Prof. Satoru Hayamizu, PhD¹, Takashi Ichinomiya, PhD², Prof. Yasutomi Kinosada, PhD²

¹Graduate School of Engineering, Gifu University, Japan; ²Biomedical Informatics, Gifu University Graduate School of Medicine, Japan

Introduction

Recently, Electronic Medical Record (EMR) systems have been widely introduced not only in hospitals but also clinics. Our goal is to extract a standard medical care pathway from numerous EMRs and compare it with a medical process obtained from EMRs in a clinic, in order to discover any managerial and medical issues in the clinic and finally reduce healthcare costs. This paper investigates a method to extract a clinical process from examination descriptions (unstructured data) in large-size EMRs applying natural language processing, and intuitively visualize the process. We also analyze and discuss the results in engineering and medical standpoints to evaluate usefulness of our proposed method. We believe our proposed process mining technique can be also applied for guiding the care process from a number of EMR data.

EMR data

We collected EMR data recorded in two clinics. There are roughly 7,100 encounters and 2,280 patients in our data. Each record consists of patient information (e.g. ID, date, age and gender), anamnesis, and free description about examination including SOAP notes. The SOAP note is a common format having subjective, objective, assessment, and plan.

Our proposed process mining method

(1) Without using the morphological analysis, all medical terms appeared in anamnesis and examination fields of each record are extracted using drug name and disease name lists. (2) A term vector indicating existence of the terms in the lists is generated. (3) Given a disease name, simple screening is applied to extract related term vectors. (4) Agglomerative hierarchical clustering is conducted to group the vectors into some classes. (5) Transitions between clusters are counted using sequential information (ID and date in records) (6) Finally, visualization is performed to show classes and transitions as a graph.

Experiment and Result

Figure 1 shows clinical processes created by our proposed method from the medical records and the history the patient has received at the clinic. We applied our scheme to two diseases: influenza (incl. common cold) and diabetes. We discussed the results subjectively based on examination fields and several criteria.

Conclusion

In this paper we proposed a method to extract a clinical process from unstructured free-form examination descriptions in EMRs. Experimental results clarified usefulness of our method. Our future work includes developing a standard medical pathway and exploring a data mining scheme to help clinic managements.

References

An Assessment Tool for Evaluating Shelters during Disasters

Masaharu Nakayama, M.D. Ph.D.¹,², Tadashi Ishii, M.D. Ph.D.²
¹International Research Institute of Disaster Science, Tohoku University, Miyagi, JAPAN; ²Tohoku University Hospital, Sendai, Miyagi, JAPAN

Abstract

After the Great East Japan Earthquake, the assessment of shelters during disasters became seen as helpful for grasping the health status of evacuees and preventing pandemics. To maintain the quality of assessment and to decrease the tedious work in inputting and analyzing enormous amounts of data, we developed a mobile assessment system with a standardized questionnaire to assess shelters during disasters.

Introduction

In March of 2011, the Great East Japan Earthquake destroyed a vast area of the northeastern coast of Japan, and as of March 2015, has resulted in 15,890 dead and 2,589 missing. During the disaster, more than 400,000 people were evacuated to temporary shelters. Continuous surveys of medical and sanitary conditions as well as environmental conditions, including power, gas, and water supplies, were successful in avoiding pandemics in well-managed shelters in the Ishinomaki area, one of the most damaged areas [1]. However, these surveys, which included calculation of the results and analysis, caused high workloads among staff.

Methods and Results

We developed a mobile assessment system using Swift 2; the application was named the Rapid Assessment System of Evacuation Center Conditions (RASSECC). The database was built using MongoDB 2.4.9, and the administrative screen interface was developed using Ruby 2.0.0p451. The application is available on iPads and iPhones with iOS 9 or over. We can store the input data in each device even offline. After the device is reconnected to the Internet, the data are sent to a server. The administrative screen is color-coded to show the severity of conditions in each shelter, and is available on web browsers such as Safari, Internet Explorer, Google Chrome, and FireFox.

Figure 1. Administrative screen

Conclusion

We developed a mobile assessment system; the input application is available on iPads and iPhones. This application may alleviate work burdens for staff teams when evaluating shelters, and the questionnaire on the application is currently being standardized among specialists of disaster medicine. Additionally, the administrative screen is available on any common browser; the table is useful for staff teams when evaluating the status of each shelter, resulting in the prompt development of strategies.

References

All Alerts are not Created Equal: A Study of Differences in User Perceptions of Drug-Drug and Drug-Allergy Interaction Alerts

Pamela Neri, MS¹, Elisabeth Burdick MS², David W. Bates MD¹, ², ⁴, Shobha Phansalkar RPh, PhD², ⁴, ⁵
¹Partners HealthCare, Wellesley, MA; ²Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA; ⁴Harvard Medical School, Boston, MA; ⁵Clinical Effectiveness, Wolters Kluwer Health, Minneapolis, MN.

Introduction Computerized Physician Order Entry systems (CPOE) with Clinical Decision Support (CDS) have the potential to reduce medication errors and improve quality of care. Studies have identified design and implementation issues with medication-related CDS alerts which can lead to provider dissatisfaction and high override rates. To understand the factors that might influence provider dissatisfaction with different types of alerts, we conducted a survey of provider perceptions.

Methods In this international, multi-site study we conducted a survey of CPOE users to assess perceptions related to receiving drug-drug and drug-allergy interaction alerts at the time of prescribing medications. These sites had systems that were both vendor-based and in-house created knowledge base systems. We distributed a previously developed and validated survey by Zheng et al.¹ to 1545 internal medicine physicians at 8 sites in the United States, United Kingdom and the Netherlands. The survey assessed provider’s attitudes about drug interaction alerts on 7 constructs: Performance Expectancy (e.g. useful in helping care for patients, clinically relevant), Effort Expectancy (e.g. reading and responding to alerts takes too much time), Social Influence (e.g. reading and responding because colleagues read and respond), Facilitating Conditions (e.g. received adequate training on how to read and respond to alerts), Perceived Fatigue (e.g. receive too many alerts), Perceived Ease of Use (e.g. find alerts easy to understand), and Perceived Use Behavior (e.g. thoroughly read alerts, provide reasons for override). For each construct of the survey, we pooled the responses to get a mean for DDI and DAI alerts. We set the completion threshold at 50% for a survey and the response rate threshold at 70% to for a construct to be included in the analysis.

Results Of the 1545 surveys distributed, 365 were opened and 48 of these were excluded due to low completion. Our analysis includes 317 surveys that were more than 50% complete. Response rates for all except one construct ranged from 88% to 99%. One construct (Social Influence) did not meet the minimum threshold for inclusion in our analysis. For five of the constructs, the difference in means between DDI and DAI alert perceptions was significant: Performance Expectancy (p<0.0001), Perceived Ease of Use (p=0.0007), Effort Expectancy (p=0.002), Perceived Fatigue (p<0.0001), and Perceived Use Behavior (p<0.0001). The differences in means for the constructs of Social Influence (p=0.42) and Facilitating Conditions (p=0.04) did not attain significance.

Discussion Survey responders answered more favorably to drug-allergy interaction alerts in comparison to drug-drug interaction alerts on all survey constructs, Previous studies on alert fatigue have described override rates overarching all types of medication-related decision support alerts. However, our study found that providers’ perceptions towards different types of alerts vary considerably. By asking providers specifically about two types of alerts, from the same system, we were able to understand these differences in their perceptions.

Conclusion Future studies should evaluate further why differences exist between drug-drug and drug-allergy interaction alerts and whether they are rooted in content, design, or clinical significance considerations specific to the type of alert. This understanding will further inform how we develop systems that change providers’ perceptions resulting in increased acceptance of alerts in EHRs.

References
Medication Self-Reconciliation Through Automated Telephony

Mark D. Newcomb Jr1,2, Kenneth D. Mandl, MD MPH1,2, Marc D. Natter, MD1,3
1Tufts University School of Medicine, Boston, MA; 2Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA; 3Harvard Medical School, Boston, MA

Introduction. It is well recognized that there are deficiencies in the quality of certain classes of research data derived from electronic health records1. Medication lists, in particular, are often incomplete, inaccurate, or outdated1,2, at least in part because patients often have medications prescribed from multiple healthcare providers and systems. We recognize that patients are often willing to directly share and augment their data for researcher usage using electronic means2, including via interactive voice response (IVR)3 and have developed an IVR application that enables subjects to accomplish patient-reported medication reconciliation using data derived from multiple clinical and research data sources. We therefore enable patients to directly indicate whether they are taking medications not yet documented, whether historically recorded medications are currently being taken, and whether their dosing schedules match the prescribed regimens.

Methods. The architecture for the application comprises 3 software components: (1) a Go application coordinating data transfer capabilities; (2) an IVR application running on an IVR engine (Voxeo CXP and Prophecy, Aspect Software Inc., Chelmsford, MA) that generates and parses VoiceXML files and (3) a Python service handling medication retrieval and mapping between NDC and RxNorm. All application components communicate via RESTful HTTP interfaces. Medication data from multiple sources, encoded in i2b2 using NDC or RxNorm, is de-duplicated by the application and reconciled by the patient utilizing voice or keypad commands over the telephone network. High-quality speech synthesis and automatic speech recognition (Nuance Communications Inc, Burlington, MA) are used for all aspects of user interaction. A mapping of National Drug Codes to RxNorm RxCUIs5 is used to unambiguously identify the medications; RxTerms6 is then used to extract relevant attributes of each medication such as strength, form, and ingredients so that these properties can be presented by the text-to-speech (TTS) engine in a logical fashion. Subjects are given an opportunity to add medications that were not present from data warehouse sources. Finally, the reconciled medication list is uploaded to i2b2 via submission of a FHIR medication list resource to a medication reconciliation cell for i2b2 that we have developed (i2me2. KM, MN, and collaborators)7.

Results. Software applications developed were capable of automatically eliciting information via IVR about current usage of medications, using data aggregated in an i2b2 data warehouse. Pilot testing suggests that our system will be convenient to use, with caller time spent for reconciliation of five medications generally lasting under five minutes. At the end of the call, data gathered during the call is uploaded directly into i2b2 with no intervention on the part of medical or research staff.

Discussion. Our system allows individual research subjects to perform self-reconciliation of medication data via IVR – a convenient mechanism that is accessible from any cellular or landline telephone, and thus available to extremely large and diverse populations, complementing other systems we have developed for web-based self-reconciliation of medications. Such approaches, which can occur without manual intervention by researchers themselves, enable ready scaling to thousands of patients with minimal incremental effort. Further, interoperability using the FHIR messaging protocol will support widespread adoption using a variety of data sources, including the i2b2 platform. Limitations of the current approach include need for tuning of TTS pronunciation of medication names and optimized recognition of drug names entered through the user’s speech. For the former, we have identified a database containing phonemes for every drug in the U.S. market, which can be used to tune our TTS engine’s pronunciations. For the latter, suggesting letter-by-letter spellings or presenting numbered pick lists under certain conditions may help ensure that medication data is entered accurately. A further limitation may lie in the recognition of various accents, but our ability to flexibly use commercial speech-to-text engines allows us to take advantage of the latest developments in the field. A next step in validation of our approach for research use is anticipated to be an exploratory study of medication self-reconciliation using IVR as compared with a web-based medication reconciliation application. This study will also include an evaluation of the absolute accuracy of our speech-to-text-based approach to medication entry.

Conclusion. The state of the art in speech recognition and synthesis software is advancing rapidly, and such systems are suitable for a diverse range of medical applications. Our system has shown that the current state-of-the-art in automated telephony is promising as a mechanism to accomplish patient self-reconciliation of medication data using standardized messaging protocols.

References

Funding: This work was supported by the National Institute of General Medical Sciences R01GM104303; and by Contract CDRN-1306-04608 from the Patient Centered Outcomes Research Institute (PCORI).
Assessing Patterns of Medication Use across VA and Non-VA Healthcare Systems

Khoa A. Nguyen, PharmD1,2; Alan J. Zillich, PharmD2; Susan Perkins, PhD3; Susan Ofner, MS3; David A. Haggstrom, MD, MAS1,4,5

1Center for Health Information and Communication, Department of Veterans Affairs (VA), Veterans Health Administration, Health Services Research and Development Service (CIN 13-416), Richard L. Roudebush VA Medical Center, Indianapolis, IN; 2College of Pharmacy, Purdue University, West Lafayette, IN; 3Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN; 4Department of Medicine, Indiana University School of Medicine, Indianapolis, IN; 5Regenstrief Institute, Indianapolis, IN

Introduction: Dual healthcare system use is a common pattern of care for many Veterans. While improving access and choice of healthcare for Veterans, it creates fragmented flows of information that can affect patient care. In an effort to reduce fragmentation, the Department of Veteran Affairs is implementing a nationwide health information exchange (HIE) program called the Virtual Lifetime Electronic Record (VLER), which allows providers to access and share information across healthcare systems. However, what data is available regarding the use of medications in regional HIEs is largely unknown. Therefore, the objective of this study is to describe the prevalence of medication dispensing across VA and non-VA health care systems prior to enrollment in VLER.

Methods: This retrospective cohort study included all Veterans who had two outpatient visits or one inpatient visit during a 1-year period prior to VLER enrollment to a local VA and HIE program. We examined outpatient dispensing during a 2-year time window prior to the VLER enrollment. Data were extracted from the VA Pharmacy Benefits Management system and medication records in a regional HIE. Source of medication data was assessed at the subject level, and categorized as VA, non-VA, or both. We then compared the mean number of prescriptions, as well as overall and pairwise differences in medication dispensing.

Results: Out of 52,444 Veterans, 17.4% of subjects (n=9,123) had medication data available in a regional HIE. On average, 39.8 prescriptions per year were prescribed for Veterans who used both sources compared to 29.3 prescriptions per year from VA only and 24.6 prescriptions per year from non-VA only sources. The annualized prescription rate of Veterans in the dual use group was 36% higher than those who had only VA data available and 61% higher than those who had only non-VA data available (Table 1).

The most common medications prescribed for non-VA source are antibiotics (17.5%), antineoplastics (15.4%), and anticoagulants (11.8%) Table 1: Annualized prescriptions by subject level source*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Overall p &lt;.0001</th>
<th>Subject Level Source</th>
<th>Incident Rate Ratio (95% CI) and adjusted p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VA</td>
<td>Both</td>
<td>Non-VA</td>
</tr>
<tr>
<td>Number of Rx per year</td>
<td>52444</td>
<td>43321</td>
<td>7424</td>
</tr>
<tr>
<td>Mean</td>
<td>30.6 ± 29.3 ±</td>
<td>39.8 ± 24.6 ±</td>
<td></td>
</tr>
<tr>
<td>± SD</td>
<td>28.2</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>Median (Min, Max)</td>
<td>23 (1, 349)</td>
<td>22 (1, 302)</td>
<td>32 (1, 349)</td>
</tr>
</tbody>
</table>

* Number of prescriptions was annualized by dividing the number of prescriptions in the two year baseline period by 2

Conclusions: Prior to enrollment in VLER, 17.4% of subjects had medication data available outside the VA in a regional HIE, including prescriptions for antibiotics, antineoplastics, and anticoagulants. This large amount of non-VA medication use is not routinely captured within VA electronic health records. Subjects receiving medication from both sources appeared to have more complex medical needs, as reflected by their higher overall mean number of medications. These data support the need for HIE programs to improve coordination of information, with the potential to reduce adverse medication interactions and improve patient safety.
Supporting and engaging Women Veterans through social media:
Informational vs. emotional support for health issues

Lien B. Nguyen, MPH¹, Priya Nambisan, PhD²
¹College of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI;
²Department of Health Informatics and Administration, University of Wisconsin-
Milwaukee, Milwaukee, WI

Introduction: Women veterans are the fastest growing group in the U.S and are projected to increase in the next
decade. There was an estimated 1.5 million women veterans in 2009, which represented 8% of the total veteran
population; this number has increased to 2.2 million in 2013.² Despite this growth, these veterans belong to an
underserved population facing unique and different challenges in a health care system originally designed to serve
mostly men. Numerous health organizations are using social media to actively engage with patients and health
consumers. A national survey reported that 87% of U.S. adults use the Internet and 72% say they have looked online
for health information.¹ People also access health information through social media sites, such as Facebook and
Twitter. However, there is limited research focused on women veterans’ health and use of social media, which is a
popular online web-based technology platform used by millions of people every day.

Objective: The study objective is to investigate what types of information is being communicated to women
veterans by supporting organizations on the social media platform, specifically on Twitter.

Methods: Seven of Twitter’s largest groups were selected for our study analysis. The groups are classified as major
supporting organizations of women veterans, and they represent a total of 51,883 followers and 31,263 posts during
2009 to 2016. Data was collected through an aggregation of tweet messages via the Twitter API (Application
Program Interface). We examined thousands of messages posted by the seven groups and then categorized them into
common themes of the information disseminated over a five-year period. Some of the disseminated information
addresses unique health issues and challenges encountered by women veteran patients today.

Results: This is a current, on-going research study. Preliminary findings show organizational outreach efforts aimed
at women veterans through use of social media to be mostly informational support (rather than emotional support).
Detailed findings will be presented at the conference.

Conclusion / Implications: The study findings will reveal how supporting organizations are using social media as a
tool for communicating key information to women veterans about health. It will provide researchers and health
professionals with a better understanding of what types of actual information are being frequently communicated to
the online community. From this, it can help guide organizational planning to focus on creating online tools for
information-seeking users among women veterans. The findings will hold important implications for future use of
social media by organizations wanting to engage with health consumers, especially underserved populations in
health care.

After participating in this session, the learner should be better able to:

1. Identify the common themes of the online information being disseminated by organizations supporting
   women veterans.

References
sheets/health-fact-sheet
   http://www.bls.gov/spotlight/2014/women-vets/
Incidence of Seven ED Visit Diagnoses are a Function of Birth Month

Scott Mankowitz, MD1, Jeffrey A. Nielson, MD, MS2
1Emergency Medicine Associates, Teaneck, NJ; 2Summa Health System, Northeast Ohio Medical University, Rootstown, OH

Introduction
Various studies have shown correlation between birth month and disease incidence. Boland et al related birth month with disease incidence1. Mining existing datasets of primary ED diagnosis and birth date might provide additional clues as to the impact of birth month on the emergency care system. This approach focuses on disease burden as a function of annual cycles. Our principle aim was to explore the relationship between birth month and emergency department visit incidence by ICD9 code.

Methods
We extracted 836,064 consecutive billing records representing 28 emergency departments in the Northeastern United States between 9/1/2014 and 8/31/2015 consisting of date of birth, date of service and primary diagnosis ICD-9-CM billing code. We excluded all visits with diagnoses that were encountered fewer than 144 times. For each diagnosis code, we compared the observed incidence with expected rates based on distribution of birthdates in the sample. The Holm approach for multiple comparisons (with p < 0.05) allowed identification of significant patterns2.

Results
After removing 62,641 records containing rare diagnoses, the sample included 773,393 ED visits representing 614 unique diagnosis codes. Chi-squared analysis using Holm p values resulted in 7 diagnoses of significance. In order, these were: Sickle cell disease NOS, Persistent vomiting, Hb-SS disease with crisis, Gastroparesis, Fetal/neonatal jaundice NOS, Chronic pancreatitis, and Acute bronchiolitis due to respiratory syncytial virus (RSV). Each of these diagnoses was statistically significant (p < 0.00000001).

Discussion
Seven ICD9 codes showed association with birth month. RSV has been shown to have a seasonal incidence3, has not been previously linked to birth month. Neonatal jaundice occurs seasonally in spring and summer. Since it is only encountered in newborns, it is unsurprising that it would be associated with birth month. The two sickle cell anemia codes likely represent the same disease. However, the annual curve of Sickle Disease NOS is sporadic (possibly representing outliers), while the curve for Hb-SS does show a clear annual distribution with a decrease in July through November. Additional studies need to be designed to understand this relationship. Also in our sample, chronic pancreatitis patients were more likely to have been born later in the year (October through December) which has not been previously demonstrated. Stratification of pancreatitis sub-type, would be helpful in a follow-up study since there may be an etiology that is more clearly seasonal.

Conclusion
Normalizing for annual birth variations, and accounting for multiple comparisons, we found that the incidence of seven ED diagnoses are a function of birth month: Sickle cell disease NOS, Persistent vomiting, Hb-SS disease with crisis, Gastroparesis, Fetal/neonatal jaundice NOS, Chronic pancreatitis, Acute bronchiolitis due to respiratory syncytial virus (RSV). Some of these have not been known to have birth date variation and should be further explored.

References
One Year Usability Results for an Electronic Respiratory Therapy Tool for the Intensive Care Unit

Matthew E. Nolan, MD, Mikhail Dziadzko, MD, PhD, Todd J. Meyer, RRT, James E. Baker, RRT, John Dyke, Ing Tong, MS, Brian W. Pickering, MB, BCh, MSc, Vitaly Herasevich, MD, PhD
Mayo Clinic, Rochester, Minnesota

Introduction

Intensive Care Unit (ICU) Respiratory Therapists (RTs) have unique information needs to best care for patients with respiratory illness, including those needing mechanical ventilation. Current systems at our institution are poorly designed for this task, with relevant information dispersed among multiple systems. AWARE (Ambient Warning and Response Evaluation) is an electronic health record dashboard that synthesizes the data needed by ICU clinicians and reduces information overload\(^1\). In order to address the unique data needs of RTs in the ICU, we developed an electronic Respiratory Therapy tool within AWARE and studied its usability.

Methods

The AWARE RT tool aggregates multiple data sources to generate a standardized report of relevant information used by RTs, including bed location, demographics, code status, ventilator/oxygenation settings, respiratory medications, airway information, invasive lines, and the respiratory problem list/plan of care (Figure 1). It can be viewed, printed or exported in different file formats. One year following open clinical deployment we conducted a web-based usability survey of all medical/surgical ICU RTs via email solicitation. We collected respondent demographic information and used a well-validated System Usability Scale (SUS) to assess overall usability\(^2\).

Figure 1. The AWARE Respiratory Therapy Report.

Results

Out of 155 invitations, 66 surveys (43%) were collected. 33% of responders were affiliated with a surgical ICU. 64% had more than 10 years of professional experience. While 100% of respondents used the tool in daily practice and 92% reported neutral/good ease-of-acquisition, median SUS score was modest at 62.5 (IQR 51-77.5). No statistically significant difference in usability was found across the length of professional experience or affiliation.

Conclusion

Although all Respiratory Therapists used the electronic dashboard tool in daily practice, overall usability was mediocre, indicating the need for revision. Observed and learned results will be used to improve the next version of the AWARE Respiratory Therapy Electronic Tool.

References

Clinical Decision Support Substantially Improved Appropriate Screening for Vitamin D Deficiency

Ryan W Novince, MD, MS1, David Bar-Shain, MD2,3, David C Kaelber, MD, PhD, MPH1,2,3
Departments of Internal Medicine1, Pediatrics2, and the Center for Clinical Informatics Research and Education3, The MetroHealth System, Case Western Reserve University, Cleveland OH

Introduction: One of the opportunities for electronic health record (EHR) clinical decision support (CDS) is to promote appropriate ordering of tests. Our EHR offers several vitamin D metabolite tests that can be ordered depending on clinical circumstance.1 These include 25-hydroxyvitamin D (calcidiol) & 1,25-dihydroxyvitamin D (calcitriol). We studied whether a point-of-ordering clinical decision support for vitamin D laboratory tests would optimize the delivery of clinical care and limit wasteful healthcare expenditures.

Methods: A point-of-ordering EHR clinical decision support advisory for vitamin D testing was developed utilizing standard clinical decision support functionality in our Epic (Verona WI) EHR. The advisory triggered when the “Vitamin D, 1,25-Dihydroxy” was selected as an independent or order-set order, in an ambulatory or inpatient setting. The modal advisory window notified the end user, “Generally, this test does NOT add value to routine screening of vitamin D deficiency. The preferred test is Vitamin D, 25-Hydroxy.” By default, accepting the advisory performed two actions for the user: (1) removed the 1,25-dihydroxyvitamin D order, and (2) added an order for 25-hydroxyvitamin D. If the user opted to continue placing an order for 1,25-dihydroxyvitamin D, the advisory required a free text comment detailing the indications. The advisory included reference hyperlinks to UpToDate resources for Vitamin D deficiency in adults, adolescents and children.2,3

Results: The vitamin D advisory was implemented system-wide on 7/2/2015. Within 90 days, it had triggered a total of 184 times. In 162 firings, the user accepted the default actions substituting a 25-hydroxyvitamin D order for a 1,25-dihydroxyvitamin D order. In 19 instances, the user proceeded with the 1,25-dihydroxyvitamin D order. In 3 instances, the user canceled placing a vitamin D test order. Overall, the electronic advisory contributed to discontinuation of 90% of 1,25-dihydroxyvitamin D tests. Documented reasons for placing the initial 1,25-dihydroxyvitamin D orders included chronic kidney disease, acute renal failure, workup for hypercalcemia, workup for renal stones, and “per consulting endocrine service.”

Discussion: The Vitamin D advisory led to reduced cost and improved efficiency. Given a $39/cost difference in the tests, the advisory saved the health system over $6,300 in unnecessary healthcare expenditures during the study period. In addition, internal data from the Clinical Core Lab show that, prior to implementation, only 16% of 1,25-dihydroxyvitamin D test orders were clinically indicated. By reducing the frequency of ordering, the advisory decreased the time laboratory personnel spent contacting providers to determine if the test was appropriate. Notably, our EHR advisory for vitamin D laboratory test orders corroborated outcomes recently reported by Krasowski and co-authors within the University of Iowa Health Care System4. Our investigation was novel in that the EHR advisory for vitamin D test orders included point of care interventions for end users.

Conclusion: This research investigation defining the impact of a novel EHR advisory for vitamin D laboratory test orders highlights the opportunity for the development and implementation of other clinically relevant point-of-orders CDS to support appropriate utilization of laboratory tests across healthcare systems.

References:
Exploring the Patient Perspective on Access, Interpretation, and Use of Test Results from Patient Portals

Daniel T. Nystrom, MS AEEP1,2, Traber Davis Giardina, PhD1,2, Jessica Baldwin, BA1,2, Dean F. Sittig, PhD3, Hardeep Singh, MD, MPH1,2

1Houston VA HSR&D Center for Innovations in Quality, Effectiveness and Safety; 2Baylor College of Medicine, Houston, TX; 3School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas

Introduction

Patient portals provide patients with online access to health information including test results, but it is unclear how patients retrieve and use test result information from patient portals. The current study examines patients’ use of online portals to access and interpret test results and identifies limitations of current patient portal designs that constrain patients’ ability to retrieve and understand their test results.

Method

Ten semi-structured interviews and three observations were conducted with patients who use a patient portal to access test results. Data from interviews and observations were transcribed and used to construct a process map that depicts the patient portal users’ perspective of the test result process. To identify steps in the test result process, comments from interviews that identify different ways patients’ access and use test result information were grouped according to the goal they hoped to accomplish with each piece of information. For example, the task step entitled “Test interpretation” was derived from different methods and sources of information patients used to help comprehend whether their test result was “good” or “bad” and to determine how this result would impact their future health.

After creating the patient-based test result process map, interviews and observations were reanalyzed to determine shortcomings of current patient portal designs that limit patients’ ability to access or understand their test results. Once relevant shortcomings were identified, they were incorporated into the test result process map to identify at which step each shortcoming occurred.

Results

The resulting process map created from interviews and observations with patient portal users portrays: the necessary task steps that need to be fulfilled for patients’ to access and interpret their test results; the methods and sources of information that were analyzed to create each task step; the general location in which each task in the test result process takes place (either in the clinic or at the patient’s home); and at which task steps shortcomings of the design of patient portals limited patients’ ability to access or interpret their test result. In total, six task steps were created and seven limitations were identified from thirty-nine interview comments and observed behaviors. Some particularly interesting shortcomings of patient portals included: 1) the lack of notification to inform the patient a test result is complete; 2) difficulty in determining the relationship between the value reported on the test result and their medical condition; and 3) the lack of ability to trend results across time.

Conclusion

Studying the patients’ perspective to depict the test result process provides a unique look into the way patients use portals to access, interpret, and inform actions regarding their health. Identifying patients’ important task goals for interpreting test results and acknowledging different sources of information used to fulfill those goals will provide direction to developing a more patient friendly design for test result interfaces. For example, one design feature uncovered by this analysis that should be incorporated into future patient portal designs is an explicit representation or description that informs patients of the relationship between the value on their test result and their medical condition (e.g., if the patient receives a low-density lipoprotein (LDL) score of 162 mg/dL, they should be informed of the optimal level for an LDL result (<100) and explained that as the LDL score increases, this is associated with a higher rate of accumulating plaque within the heart’s arteries; resulting in an increased chance of developing cardiovascular disease.) Current and future work will continue to use this diagram to develop a test result interface that enhances patients’ ability to access, comprehend, and follow-up on test results obtained from patient portals.
Purpose: Systemic lupus erythematosus leads to renal disease (lupus nephritis (LN)) in half of patients, many of whom progress to renal failure. The American College of Rheumatology guidelines for the treatment of LN recommend monitoring of all SLE patients for incident LN. If LN occurs, a renal biopsy is recommended to characterize disease. At the time of LN diagnosis, therapy should be started with concurrent laboratory monitoring of therapeutic response. If improvement is not seen in the prescribed timeframe, the provider should switch or add additional therapies. Guideline adherence must be timely, as care delays can lead to renal failure. Guidelines steps are often missed due to patient, provider, and system factors. Novel predictive analytics tools could identify patients at risk for treatment failure so that personalized interventions can be instituted before irreversible renal damage occurs. The goal of this project was to design a LN clinical outcome prediction tool that identifies incident LN cases and reports a treatment failure risk score in the medical record. This implementation would allow limited care coordinator resources to provide sustainable, consistent, and efficient adherence to guidelines in at risk patients.

Informatics Solution: As advocated by the Agency for Healthcare Research and Quality, our intervention is based on a conceptual framework. The conceptual model includes LN antecedents such as: an individual’s personal characteristics (including race/ethnicity, age, gender, body mass index, and smoking; patient characteristics e.g. co-morbid conditions such as diabetes, hypertension, and depression) and LN characteristics (such as duration of disease and severity of disease by measuring baseline estimated glomerular filtration rate (eGFR), urine protein/creatinine (UPrCr), C3, C4, anti-double stranded DNA (anti-dsDNA) antibodies, blood pressure, and classification of renal lesions by biopsy). The outcome risk score predicts treatment response probability using ACR Guideline outcomes (eGFR, UPrCr). Our group developed a random forest model based on a patient’s baseline LN characteristic labs to predict complete response to therapy at one year. This model’s receiver operating characteristics area under the curve for outcome was 0.74. During Phase 1 of the project, this lab-based model will calculate the risk score for LN patients. During Phase 2, model refinement using the personal, patient, and disease derived antecedents and longitudinal LN characteristics described above will be captured and utilized using over 20 years of data from the Enterprise Data Warehouse (EDW). This includes data such as pathology reports (annotated by NLP), billing diagnosis codes, and laboratory values. NLP annotation of text progress notes will be used to populate vital sign data from a legacy medical record. Random forest modeling will be used to refine the model, and the risk score from the refined model will be reported as on Phase 1. The data flow scheme for both phases will occur as in Figure 1. InfoSphere® BigInsights from IBM® is utilized for storing and analyzing large volumes of both structured data (from the EDW, which stores discrete laboratory and diagnosis code data), and unstructured data (annotated by Natural Language Processing (NLP), first using the Linguamatics®’ algorithms, followed by BigInsights processing). BigInsights will interface with SPSS for random forest model calculation of a risk score to populate an Epic smart data element. A registry of LN patients (defined by ACR guidelines) with a complete response risk score as a reporting metric will be created and used to populate a care coordinator dashboard with high, medium, and low risk patients (based on the risk score) to prioritize patient and provider contact to optimize care.

Conclusion: Novel predictive modeling risk scores for LN outcome can be presented in the medical record so that valuable care coordination efforts can be triaged to patients most at risk for treatment failure.
Carelign®: Early Reactions to a Mobile Care Management and Handoff Tool

Jillian Olsen, MD1, Jacqueline Soegaard, SB1, Geoffrey Bass, MD, MBA2, Ryan Leone3, Glenn Falai2, Rich Urbani2, William Hanson III, MD1,2, Subha Airan-Javia, MD1,2
1Perelman School of Medicine, Philadelphia, PA; 2University if Pennsylvania Health System, Philadelphia, PA; 3University of Pennsylvania, Philadelphia, PA

Introduction

Failures in transitions of care are a well-established risk to patient safety. Written handoffs can reduce that risk by ensuring the constant availability of accurate information. Our legacy handoff tool, used at our institution through 2015, consists of an array of text boxes that were used in diverse ways and updated irregularly, limiting readers’ ability to trust the information. Carelign® is an internally-developed responsive web application that imports real-time patient data into a handoff platform accessible and editable from all electronic devices, where all providers manage a common care plan. We hypothesized that this interactive handoff with up-to-date clinical data, a shared care plan, and mobile access would improve (1) inter-provider communication, (2) efficiency, and (3) satisfaction.

Methods and Results

In early 2016, we compared provider perceptions of our new handoff tool (Carelign®) and the previous handoff tool (PHT). The general medicine services at the Hospital of the University of Pennsylvania are naturally divided into 2 groups with identical structures and nearly equal numbers of patients; 1 group continued to use the PHT while the other transitioned to Carelign®. After 1 to 8 weeks, 51 physicians responded to 16 Likert-type scale statements (Figure 1). The percent responding positively (“Agree/Strongly Agree”) and negatively (“Disagree/“Strongly Disagree”) were compared via a Chi-squared test, with statistical significance in favor of Carelign® found for 3 statements: “The current written handoff system is safe for patients” (p = 0.022), “I am satisfied with the current handoff system” (p = 0.012), and “I would recommend the current handoff system to colleagues” (p = 0.011). Responses among early Carelign® users (weeks 1-4) and late Carelign® users (weeks 5-8) were compared to each other in the same fashion, with no statistically significant differences found between the 2 groups.

Figure 1. Percent of providers using each system who responded positively (“Agree”/“Strongly Agree”).

To assess Carelign’s® effect on written communication, patients’ “summary statements” were evaluated for whether they had been updated daily. A total of 730 statements were examined, revealing that 24.7% of Carelign® and 21.7% PHT statements were newly edited and thus presumably more current. Additionally, 152 patients cared for by these teams were asked in person to respond to Likert-type scale statements about provider communication and perspectives on healthcare IT. Notably, 72% replied positively to, “My care providers should use mobile devices to access and manage my health information,” thereby negating one anticipated concern about a mobile platform.

Conclusion

These early data suggest that changing the handoff tool to one with an intuitive user interface with real time patient information, mobile accessibility, minimal training requirements, and standardized data entry can improve perceptions of communication, efficiency, satisfaction, and patient safety. Objective measures of the reliability of written handoffs suggest a positive trend, and patients view clinicians’ use of mobile devices favorably. Studies of Carelign’s® impact are ongoing and will have broad implications for future improvements in transitions of care.

© Perelman School of Medicine, University of Pennsylvania, PA

1539
Application of Ontobedia on Cell Line Ontology Development

Edison Ong¹, Yongqun He, DVM, PhD¹
¹ University of Michigan Medical School, Ann Arbor, MI, USA

Introduction
Ontology has played an important role in representing biomedical knowledge and data. The community-based Cell Line Ontology (CLO) records and annotates over 35,000 cell lines¹. It is difficult to edit and annotate so many cell lines without community involvement. Ontobedia, a Wikipedia-like knowledgebase with ontology editing features, is applied to support the collaborative editing and distribution of CLO². This report presents the usage of Ontobedia and its enhancement on the community-based development of CLO.

Ontobedia Usage for Community-based CLO Development
Ontobedia is developed based on MediaWiki software platform with Ontokiwi extension support². MediaWiki is the software that runs Wikipedia server. Ontokiwi is an extension to MediaWiki that supports ontology query, storage and modification. The MySQL relational database and Virtuoso triple store database are used for both Wikipage and ontological data access, storage, and processing. The latest version of CLO is loaded to Ontobedia using Ontokiwi importing function. The process of community-based development of CLO through Ontobedia web server is illustrated (See example, Hela cell: http://ontobedia.hegroup.org/index.php/CLO:CLO_0003684).

1) WikiText Editor: Ontobedia provides convenient access for users to edit existing cell line information via Wikitext source editor. The extended Wikitext supports ontology editing with some specifically designed formats on top of the original syntax. For example, we use the text “label =”, “subclassof =” and “equivalent =” to represent the type of information and the text following “=” sign is automatically validated and re-formatted appropriately. A cell line can also contain text descriptions that are not linked to the ontology annotations and axiom definitions. Users can edit this information using the original MediaWiki Wikitext format similar to Wikipedia.

2) Ontology Form Editor: A cell line can also be annotated via the form-based editing system, which contains four major ontology editing sections: (i) General information includes the label IRI and type of the cell line. Note that IRI and type are restricted to the administrative user because improper modification may introduce ontological conflicts. (ii) The hierarchical definition of each cell line can be easily entered with the help of cell line searching function. (iii) Annotation information can be modified with pre-defined annotation properties such as “editor”, “comment” and “version”. (iv) Axiom definition of the individual cell line can be defined via W3C Manchester format.

3) Version Control and Discussion: Wiki page version control and discussion page are two important features in MediaWiki and such features are naturally inherited in Ontobedia. As a result, Ontobedia supports history checking, version control and social discussion. After we make text changes, the history of the changes is recorded and the information can be displayed in the future. A user can compare the text differences at two selected time points. Users can also raise their concerns or questions on the discussion page of the individual cell line.

A more comprehensive tutorial is located at http://ontobedia.hegroup.org/index.php/Ontobedia_tutorial.

Conclusion
Ontobedia has many unique features to support community ontology development compared to other existing tools such as WebProtégé (http://webprotege.stanford.edu/) and Semantic MediaWiki (SMW; https://www.semantic-mediawiki.org). WebProtégé does not support the Wikipedia style non-ontology text editing, and its usage requires professional ontology training. While SMW extends Wikipedia features, SMW treats normal pages as instances instead of ontology classes, and therefore it is not designed to support ontology development.

Acknowledgement: The project is funded by a MCubed Diamond program in the University of Michigan.

References
Informatics in the International Medical School Curriculum: A Pilot Survey

Amy Opalek, M.S.¹,², Sarah W. Leng, M.A.¹, Danette W. McKinley, Ph.D.¹
¹ Foundation for Advancement of International Medical Education and Research (FAIMER), Philadelphia, PA; ² PhD Student, Information Studies, Drexel University College of Computing and Informatics, Philadelphia, PA

Background
A recent meta-analysis has shown the importance of IT competencies and experience among clinical leaders. As clinical leaders often come from the medical profession, the extent to which informatics is addressed in the medical curriculum is a matter for concern. Unfortunately, another recent study has shown that, despite their interest in clinical informatics, medical students are often unaware of available training opportunities in the field. Further, data and research describing the clinical informatics workforce outside developed, English-speaking countries is limited. This pilot study aims to describe the extent to which international medical schools report having biomedical informatics (BI) in their curricula, and any variations that may occur by region or school type.

Methods
A questionnaire was distributed to 227 medical schools outside the United States and Canada that participate in an ongoing data exchange program with the Educational Commission for Foreign Medical Graduates and its foundation, FAIMER. The questionnaire included items about the schools' programs, including details about required and elective BI courses within the medical program and independent BI degree programs offered. Medical school demographic data from the World Directory of Medical Schools, including school type (public/private) and location, was merged with the questionnaire response data. A chi-square test of statistical significance (SPSS v. 22) was used to assess the relationship between school funding type and BI offerings.

Results
117 (51.5%) questionnaire responses were received from schools in all world regions, with more than five surveys each from the United Kingdom (8), Australia (7), and Japan (7), and Poland (7). Over half (n=66; 57.9%) of responding schools were private medical schools. Of the 113 who answered the question, only 21 (18.6%) reported offering a BI program at their school. Only one of the private schools offered such a program; the relationship between school type and offering a BI program was statistically significant (p<.01). Within the medical curriculum, 40 of 112 respondents (35.7%) reported having a BI course, with 28 (25.0%) schools indicating this was a required course rather than an elective. An additional 32 respondents indicated that the topic was covered across other courses in the curriculum. Among responding Australian schools, 6 (85.7%) offered BI within the medical curriculum, while only 2 (25%) in the United Kingdom did so.

Conclusion
There is wide variation in the extent to which BI education is available in medical schools around the world. Full BI programs appear more likely to be available at public rather than private medical schools. However, a minority of medical schools surveyed offer full programs in BI, and, worryingly, over a third indicated no BI exposure in their medical curriculum. There appears to be much room for improvement internationally in the education of future clinical leaders in informatics issues. Further studies should sample schools more broadly and longitudinally to attempt to ascertain whether there are regional differences in the coverage of BI issues and whether a greater number of schools choose to include BI in their curricula over time.

References
Pediatric Inpatient Medication Reconciliation using EMR in a Community Hospital

Haifa Jaedi, MD1, Chionye Ossai, MD1, Fernanda Kupferman, MD1, Diana Aschettino, MD1

1Brookdale University Hospital and Medical Center, Brooklyn, New York

Introduction

According to the Joint Commission, medication reconciliation is the process of comparing a patient’s medication orders to all the medications that the patient is currently taking. Reconciliation is done to avoid medication errors and EMRs provide a good medium for the reconciliation process. This study was a quality improvement project aimed at increasing the medication reconciliation rates among patients admitted to the pediatric inpatient/ICU departments of Brookdale University Hospital and Medical Center.

Methods

A retrospective chart review of all patients admitted to the pediatric department in June and July 2015 was done and baseline medication reconciliation rates were noted.

In the first phase of intervention, computer generated prompts and alerts were designed to remind residents about reconciling patient medications upon admission. Alerts were generated every time the patient’s chart was reviewed without a complete reconciliation process. Residents and faculty were formally trained on the importance and process of medication reconciliation using the electronic medical record.

The second phase of intervention required senior residents to manually perform spot checks for adherence to the medication reconciliation process by reviewing H&P notes and reconciling patient-reported medications with medications actually entered into the electronic medical record.

Results

![Fig 1: Medication Reconciliation Rates](image)

Discussion

We were able to accomplish our goal of 90% or more reconciliation rates with our interventions. The graph above reports the rates following our second intervention. The rates were much higher (up to 98% in January 2016) without the manual reconciliation by senior residents because some patient-reported medications in the H&P notes were not entered into the EMR. EMRs therefore might have some limitations in the medication reconciliation process. Further studies need to be carried out to assess the quality of the reconciliation process using EMR generated prompts and alerts.
Using Tablet Computer Journals to Capture Patients’ Chronic Disease Management in Daily-living Settings

Mustafa Ozkaynak, PhD¹, Katia Hannah, MPH¹, Gina Woodhouse, BS², Rupa Valdez, PhD³, Patrick Klem, PhD²

¹University of Colorado Anschutz Medical Campus, Aurora, CO; ²University of Colorado Hospital, Aurora, CO; ³University of Virginia, Charlottesville, VA

Abstract

We tested a tablet computer journal designed to capture both health activities and relevant challenges and facilitators in daily-living settings. Our feasibility study showed that tablet computer journals could capture health-related routines and incidents affecting anticoagulation management. Our approach informs consumer informatics interventions by revealing individual needs and social/organizational contexts. This approach also informs collaborative informatics interventions by revealing information related to chronic disease management in daily-living settings that should be discussed with clinicians.

Introduction

Individuals with chronic conditions are often required to perform various health-related activities in daily-living (home and public) settings. Examining these activities, including the challenges and facilitators patients face while performing these activities, can inform consumer informatics interventions. Traditional workflow methods such as observations have been used effectively to capture activities, including relevant challenges and facilitators, within well-structured clinical settings. However, these methods may be more difficult to apply in the unstructured daily-living settings of patients’ homes and communities. Informatics researchers and designers need new methodologies to understand chronic health management in this context. The purpose of this study is to test a tablet computer journal designed to capture health activities as well as relevant challenges and facilitators in daily-living settings.

Methods

From October 2015 to March 2016, thirteen patients receiving anticoagulation pharmacologic therapy were recruited from the anticoagulation clinic at a Western academic hospital. Patients were given a tablet computer with a customized 11-question survey to journal about daily-living health activities pertaining to anticoagulation therapy management for one month. The tablet computers included voice entry software. Eleven patients (undergoing anticoagulation therapy between 1.5 months and 25 years) participated in the study until completion. Patients were predominantly female, aged from 26 to 71 years. After journaling, participants were interviewed about tablet usability and journal-entry content. The Colorado Multiple Institutional Review Board approved this study.

Results

One hundred twenty-nine tablet entries were made during this five-month timeframe. Early analysis of the journals revealed patient frustration with bruising and bleeding, impact of social environment on treatment, difficulties with food restrictions, details of medication use (e.g., time of day, use of tools), and situations where they failed to take medications. Feedback pertaining to tablet computer usability in tracking daily-living health activities indicated that tablets were of appropriate size and application formatting was ideal. Patients also found voice entry helpful. Technical issues encountered were mainly a result of poor Internet connection.

Conclusion

Our preliminary study showed the feasibility of tablet computer journals in capturing health-related routines and incidents that affect anticoagulation management. Our approach also demonstrated feasibility in capturing some challenges patients faced at home. One strength of journaling is that data are collected in near-real-time, when the memory of the participant is fresh. Journal information also facilitated obtaining rich information during follow-up interviews. Important limitations of this method could be cost and the need for administrative support. The approach we tested can inform consumer informatics interventions by revealing individual needs and social and organizational contexts. This approach can also inform collaborative informatics interventions by revealing important information related to chronic disease management in daily-living settings that should be discussed with clinicians. This study was supported by AHRQ under award number R03HS024092 (PI: Ozkaynak).
Visualizing Clinical Event Sequences to Support EHR Data Retrieval

Brian M. Paciotti, Ph.D., M.S.¹, Fan Du, M.S.², Shang Wei, M.S.¹, Samuel Morley, M.S.¹
¹University of California, Davis Health System, Sacramento, CA; ²Human-Computer Interaction Lab, University of Maryland, College Park, MD

Abstract

EventFlow is an application that helps analysts visualize complex temporal patterns—often an important task when retrieving and processing clinical data for retrospective observational studies. EventFlow illuminates how event types and sequences impact particular research questions. The visualization process improves the efficiency of data retrieval by revealing more specific requirements before analysts have to create complex query logic. The application also highlights data quality issues such as duplicate records or 100 year old babies.

Visualizing Temporal Sequences with EventFlow

The Human-Computer Interaction lab at the University of Maryland developed EventFlow to help analysts visualize categorical temporal event data. It has been applied to domains such as cybersecurity, sports analytics, and healthcare management. The application creators believe that user interfaces are evolving toward larger, information-abundant interactive visual displays that will help analysts assess data quality, compare populations, and spot actionable anomalies. Figure 1 illustrates the three panels on the EventFlow interface: the left panel has controls and a legend, the middle panel summarizes all sequence patterns in the dataset, and the right panel has a scrollable timeline browser showing all of the individual records. The top sequence in the overview is selected (drug A, followed by stroke, followed by drug B). The distance between events corresponds to the average time between events. The height of the bar corresponds to the proportion of records with that sequence.

Figure 1. EventFlow with a small sample dataset.

Using EventFlow to Improve EHR Data Retrieval Requests

Researchers are increasingly asking for EHR data extractions for observational research and many of these studies focus on the timing and sequence of clinical events. Analysts have been using EventFlow to communicate complex event sequences with researchers and facilitate a more efficient way to gather requirements and implement SQL query logic. EventFlow has improved the data retrieval process by reducing the time required to solidify requirements and separate data that are true signal to the analysis (e.g., specific types of events) from the noise (e.g., duplicates, too many rows of redundant clinical events, data errors).

References

Sociotechnical Interventions to Optimize Research-Focused EHR Data Retrieval Requests

Brian M. Paciotti, Ph.D., M.S., Shang Wei, M.S., Samuel Morley M.S.
University of California, Davis Health System, Sacramento, CA

Abstract
An increasing number of researchers are requiring high-quality and pre-processed clinical EHR data for observational research and cohort discovery. To meet these challenges, a grant-sponsored data retrieval unit is implementing social and technical interventions to improve the quality and speed in which processed datasets are delivered to researchers. Technical interventions include more advanced datamarts and social interventions involve additional training and more strategic communication with clinicians.

Summary of Current Data Retrieval
The University of California Davis Health System (UCDHS) implemented an EHR supported by Epic in 2003. Since about 2006, various units at UCDHS have been responsible to create reports and extract data from the EHR. An informatics team supported by a Clinical and Translational Science Award (CTSA) has been responsible to retrieve data for research projects (e.g., cohort discovery for patient recruitment, retrospective observational studies). This small team is tasked with processing requests, defining requirements through in-depth meetings, and retrieving data from the EHR. When time allows, the team has worked hard to format the data into readily analyzable files. Recently the team implemented a requirements screening questionnaire to better assess early on what particular data specific researchers require. A new era has arrived. Organizational leaders have changed their driving questions from, “Can informatics teams extract clinical data?” to “How can informatics professionals create more efficient sociotechnical systems to provide a larger number of researchers with high-quality data?”

Process Improvement
The UCDHS informatics team is pursuing a number of technical and social interventions to improve the data retrieval process. From a technical perspective, the team is improving the effectiveness of the I2B2 “Cohort Discovery” tool to better allow researchers to query and visualize de-identified output. For example, researchers are better able to refine their requirements when they have the ability to create their own complex queries and visualize important traits of the cohort such as age, gender, and race. In addition, analytical datamarts supported by a standards-based informatics platform are being created so that analysts can more efficiently extract data. The team has investigated a number of social interventions: clinical champions/gatekeepers to bridge communication between clinicians and analysts, documentation of the retrieval process, and training sessions. A specific social intervention is to better inform clinical researchers about five EHR data quality dimensions—completeness, correctness, concordance, plausibility and currency.²

Conclusion
The growth of clinical data within EHRs combined with a increasing interest in observational research will continue to pressure informatics teams to improve their data retrieval processes. The UCDHS informatics team is facing these challenges by simultaneously improving both technical and social aspects of the data retrieval process.

References
Text Processing of Clinical Research Protocols and Informed Consents to Facilitate Tracking of Research Procedures

Alok Sagar Panny BDS¹, ², Vojtech Huser MD PhD²
¹Department of Health Administration and Policy, George Mason University, Fairfax, VA
²Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction: Clinical Research Informatics (CRI) aims to improve the conduct of clinical research studies. Research decision support (RDS) alerts and reminders can monitor whether study protocol is being followed. This support is similar to clinical decision support (CDS) systems that monitor compliance to clinical guidelines. A necessary pre-requisite for provision of research decision support is a computable representation of a study protocol.

Background: National Institutes of Health Clinical Center (NIH CC) conducts a large number of clinical trials (in 2014, it has 1611 active research protocols, of which 48% were interventional (phase 1-4), 46% natural history and 6% other). NIH CC maintains a research data warehouse and numerous clinical research informatics systems, including a system to manage study protocols and informed consents. From an informatics perspective, computable representation of study milestones and study procedures for every research protocol could be beneficial to numerous research system, including electronic data capture (EDC) systems, electronic health record (EHR) system or research cost accounting. However, with a large number of protocols and different needs of various CRI systems, the current CRI software relies mostly on traditional formats (such as semi-structured text or PDF) to store and manage clinical study metadata and study documents (specifically the study protocol and study informed consent [IC]). Computable representation of research studies is an ongoing CRI challenge with only limited support by existing CRI standards.

Rather than requiring researchers to use additional protocol modelling software, our study pilots an approach where we use natural language processing (NLP) methods on study documents (protocol or IC) to detect study procedures. Extracted study milestone events (or procedures) can be subsequently used to track progress of individual study participants through the protocol by observing the events recorded in the research data warehouse.

Methods: We created a pipeline of multiple text processing and NLP tools that starts with protocol or informed consent as input. For PDF documents, we first extract the text and remove repeating header and footer text. The current pipeline uses MetaMap as the main NLP tool but we are evaluating several other tools (e.g., Noble Coder, Apache cTAKES). To assess the quality of the procedure terms NLP extraction, we created an evaluation reference standard (for a random subset of protocol documents). Examples of protocol procedures targeted in our pilot are: ‘whole blood count test’, ‘liver biopsy’ or ‘questionnaire administration’. We mark by special flag procedures that are traceable via the NIH CC data warehouse.

Preliminary Results and Discussion: From NIH CC, we obtained 2,013 ICs (all PDF files) originating from 764 active research studies (some studies had multiple versions of ICs). In addition to ICs, we also obtained 3 full protocols and 21 protocol synopsis from inttrialshare.org data sharing platform. The evaluation gold standard data are available at https://dx.doi.org/10.6084/m9.figshare.3100765.v2 (the link also contains additional results). We wrote scripts (in R language) that invoke MetaMap remote API, parse MetaMap text (or XML) output and filter only procedural detected concepts. Our comparison of multiple MetaMap configurations (restricting by semantic type [e.g., only procedures] or restricting by UMLS terminology [e.g., only SNOMEDCT_US] indicate best results with the least restrictive NLP configuration and more intensive post-processing of MetaMap outputs. Our results indicate the need for better representation of protocol documents (PDF format disadvantages are loss of formal document sections (heading and subheadings), rich text formatting and tables formatting). We have created a proposed extension to the CDISC Operational Data Model (ODM) and plan to collaborate closely with CDISC XML technology team to advance existing standards towards a computable representation.

References

Annotating Patient’s Smoking Status From Electronic Dental Record Histories

Jay Patel, BDS, MS¹, Zasim Siddiqui, BDS, MS¹, Danielle Mowery, PhD², Thankam Paul Thyvalikakath, DMD, MDS, PhD¹,³
¹Indiana University School of Dentistry, Indianapolis, IN; ²Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; ³Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN

Introduction
Smoking is a major risk factor for periodontal disease and can be leveraged to assess a patient’s dental health status over time. However, a patient’s smoking status is typically documented in free-text histories within the electronic dental record (EDR); therefore, manually abstracted for use within periodontal risk prediction models. Natural language processing (NLP) can automatically abstract patient smoking status from these histories eliminating human effort. The long-term objective of this study is to develop an NLP system to extract patients’ smoking status from the free-text EDR histories for study within a periodontal risk prediction model. In this study, we report our smoking schema and our reliability applying the schema to EDR patient histories.

Methodology
In this pilot study, we randomly sampled 100 patients and their smoking histories (December 31, 2011-January 1, 2012). First, we developed a schema representing the types of patient smoking statuses described within patients’ medical histories by conducting a literature review of smoking status definitions (top-down modeling). Simultaneously, we manually reviewed 100 patients’ (5 batches of 20 patient medical histories) to refine our smoking status subtypes (bottom-up modeling). Two trained annotators independently annotated each batch of patient histories. Any disagreements between the annotators were discussed and agreement was reached using consensus. We report our inter-annotator agreement (IAA) (F1-measure) between annotators for each batch both overall and across smoking status subtypes.

Results
We observed an overall IAA of equal or greater than 85% across batches. Intermittent, heavy, and past smoker descriptions were not always observed (NOF) in each batch.

Table 1. Smoking status schema subtypes with IAA between two annotators over 5 batches.

<table>
<thead>
<tr>
<th>Description</th>
<th>Batch1</th>
<th>Batch2</th>
<th>Batch3</th>
<th>Batch4</th>
<th>Batch5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-smoker</td>
<td>never smoked tobacco</td>
<td>98%</td>
<td>80%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Past smoker</td>
<td>smoked in the past</td>
<td>100%</td>
<td>71%</td>
<td>100%</td>
<td>NOF</td>
</tr>
<tr>
<td>Current smoker</td>
<td>smokes at present</td>
<td>0%</td>
<td>100%</td>
<td>89%</td>
<td>100%</td>
</tr>
<tr>
<td>Intermittent smoker</td>
<td>smokes occasionally</td>
<td>67%</td>
<td>NOF</td>
<td>NOF</td>
<td>NOF</td>
</tr>
<tr>
<td>Light smoker</td>
<td>smoked &lt; 1 pack per day</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Heavy smoker</td>
<td>smokes &gt; 1 pack a day</td>
<td>80%</td>
<td>100%</td>
<td>NOF</td>
<td>NOF</td>
</tr>
<tr>
<td>Overall</td>
<td>95%</td>
<td>85%</td>
<td>99%</td>
<td>100%</td>
<td>96%</td>
</tr>
</tbody>
</table>

NOF=no observed finding

Conclusions
We achieved high reliability annotating patient smoking status from dental medical histories. In future work, we will leverage this dataset to develop and evaluate an NLP tool that provides a patient’s smoking status to a periodontal risk prediction model.

Reference
Implementing a course in computational thinking during residency using R

Furquan Pathan, MD1, Sary Beidas, MD, MBI, FACP, PMP1
1GME Orange Park Medical Center, Orange Park, FL

Introduction
A widening chasm exists between what future healthcare providers are learning and the actual health care environment needs.[1,2,3] HITECH Act was introduced in 2010 to facilitate adoption of electronic health records, since then the majority if not all residency programs have adopted clinical information systems (CIS).[4] Hence, new skills and competencies are needed if healthcare providers are to effectively leverage these technologies for patient care, education and research. In particular, there is a gap in statistical analysis, visualization of data skills, and computational thinking. R is an open source application that is widely used in academia and addresses the identified gaps in education pertaining to statistics, visualization science, and computational thinking. [5,6] Because ACGME requires residents and faculty to participate in scholarly activities, we developed an introductory curriculum for R in an internal medicine residency.[7]

Methods:
The curriculum is administered using a “flipped” classroom method to residents during at the PGY-I level. In brief, the intervention is administered through prerecorded videos (8-10 minutes), classroom lectures (1 hour each) demonstrating specific skills in using R to analyze datasets, and one-on-one sessions during each resident’s research rotation, where an instructor helps the resident solve problems in a clinical data set in R. To measure the effectiveness of the course content and delivery, we administered a pre and post course survey to assess their baseline level of understanding and also to receive feedback.

Results:
Three out of ten residents have completed the introductory portion of the course. The overall reception of the course was positive in the pre-course survey, with learners showing a high level of motivation to continue expanding their knowledge in R for use in clinical problem solving and decision support. The post-course survey noted that the residents found the course moderately difficult, with no previous exposure to programming or computational thinking. A major challenge for the residents was interacting with the command line interface especially adapting to the new syntax to execute functions.

Conclusion:
Residents have the capacity and are eager to learn computational skills. The clinical curriculum can be modified to accommodate an introductory course in statistics, data, and visualization to achieve these goals. Future plans will focus on developing and personalizing course content to support resident education and needs.

References
Cues for PE Diagnosis in the Emergency Department: A Sociotechnical Systems Approach for Clinical Decision Support

Brian W. Patterson\textsuperscript{1,2}, MD MPH, Erkin Otles\textsuperscript{1}, MS, Ann Schoofs Hundt\textsuperscript{2}, PhD, Peter Hoonakker\textsuperscript{2}, PhD, Shashank Ravi\textsuperscript{1}, MD, and Pascale Carayon\textsuperscript{2,3}, PhD

\textsuperscript{1} Department of Emergency Medicine, University of Wisconsin School of Medicine and Public Health
\textsuperscript{2} Center for Quality and Productivity Improvement, University of Wisconsin-Madison
\textsuperscript{3} Department of Industrial and Systems Engineering, University of Wisconsin-Madison

Pulmonary embolus (PE) is among the most challenging diagnoses made in the emergency department (ED). While missed or delayed diagnosis of PE is a major problem in the ED\textsuperscript{1}, overtreatment, which subjects patients to harm from radiation, overdiagnosis, and increased cost, is also a concern\textsuperscript{2}. Health information technology, such as clinical decision support, has the potential to reduce diagnostic errors and support the diagnostic process.\textsuperscript{3} However, this requires that the technology be useful and usable, and fit within the clinical workflow, providing justification for a sociotechnical systems approach.\textsuperscript{4} The purpose of this study is to understand cues in the PE diagnosis process in the ED sociotechnical system and to compare these cues to the information available in the EHR. This will help in defining design requirements for a clinical decision support for PE diagnosis in the ED.

Using the Critical Decision Method,\textsuperscript{5} we interviewed 16 attending physicians and residents in three EDs of two academic medical centers and one community hospital. The total duration of the interviews was over 12 hours. Using an iterative qualitative content analysis, we identified 4 categories of cues: (1) patient signs and symptoms (e.g., leg swelling, chest pain), (2) patient risk factors (e.g., immobilization, surgery or trauma, cancer), (3) explicit risk scoring (e.g., PERC), and (4) clinical judgment. We then mapped these cues to information available in the EHR at one of the participating hospitals. About 80-90\% of the cues may be available in the EHR; many of them rely on the physical exam and information obtained by talking to the patient. This finding underlines the need to identify the various roles involved in obtaining, documenting and reviewing the information that informs the PE diagnostic process. The PE diagnostic process in the ED is distributed across multiple roles, individuals and technologies in a sometimes chaotic and often busy physical and organizational environment. Our findings highlight the potential for decision support tools to aid providers by collecting data already available in the EHR in disparate areas of the workflow and highlighting relevant cues for providers at the time of diagnostic decision making.

Acknowledgments
This project was supported by grant number R01HS022086 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

References

Evaluation of UMLS Term Coverage for Echocardiogram Measures

Olga V. Patterson, PhD1,2, Matthew S. Freiberg, MD, MSc3, Cynthia A. Brandt, MD, MPH4,5, Scott L. DuVall, PhD 1,2
1VA Salt Lake City Health Care System; 2University of Utah, Salt Lake City, UT, USA; 3Vanderbilt University Medical Center, Nashville, TN, USA; 4Connecticut VA Healthcare System; 5Yale University, West Haven, CT, USA

Abstract
The Unified Medical Language System (UMLS) is a widely used resource for concept identification systems that link phrases found in text to the concept entries in the UMLS. Building a natural language processing (NLP) system for the purposes of echocardiogram measures extraction from unstructured clinical notes requires a comprehensive lexicon. Our previous effort resulted in a custom lexicon that contained 408 entries for 27 concepts typically found in echocardiogram reports. Using interactive MetaMap, we determined UMLS coverage of the terms in the lexicon by searching for the lexicon entries in UMLS. Overall term coverage was measured at 43.9%.

Background
Since the first days of inception of the Unified Medical Language System (UMLS) multiple efforts have been performed to determine the limits of UMLS term coverage.1,2 These studies agree that while the UMLS concept coverage varies across clinical subdomains, it generally higher than the term coverage. The U.S. Department of Veterans Affairs Corporate Data Warehouse (CDW) is a national data repository comprising data from several Veterans Health Administration (VHA) clinical and administrative systems. With a large number of facilities within VHA, the unstructured data in CDW shows a high level of variability across different facilities and time. As a part of a larger NLP system development project, we built a custom lexicon to be used for term identification of echocardiogram measures in clinical notes.3

Method and Results
Using a previously described term extraction method, we obtained a custom lexicon containing 502 terms for 74 concepts. The project focused on 27 concepts decided by domain experts as the target concepts for the study, which combined had 408 terms. Each of these terms was searched in UMLS version 2013 using interactive MetaMap application offered by the National Library of Medicine. We did not use UMLS Metathesaurus Browser because MetaMap offers phase segmentation, word disambiguation, and filtering functionalities that minimize the number of ambiguous results as well as enable composite concept matches. For each of the 27 concepts in our custom lexicon, we counted the number of terms that had 1) a single mapping with a) an exact match; b) a composite concept match; 2) multiple mappings with a) at least one exact match; b) at least one composite concept match; 3) a) one or more mappings with all of them being incorrect mappings; or b) no mappings.

Out of 408 terms, only 54 terms (13.2%) had an unambiguous match to a simple or composite concept in the UMLS. Another 125 terms (30.6%) were ambiguously mapped to multiple concepts but at least one of them matched the targeted concept. 229 terms (56.1%) were either mapped to unrelated concepts or were not mapped to any concepts.

Discussion
UMLS is frequently used as a single source of terms for concept identification in NLP systems. Our study indicates that UMLS provides only 43.9% coverage of terms used in the VHA clinical notes to describe echocardiogram measures. In order for an NLP system to achieve a high level of recall in concept identification of echocardiogram measures, a custom lexicon derived from the data is required. While the UMLS contains at least one entry for each of the targeted concepts (simple or composite), the synonym set lacks all spelling variations found in clinical notes.

Acknowledgements
This work was supported by HL095136 grant from the National Heart, Lung, and Blood Institute and using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204.

References
NDC Properties in RxNorm

Lee Peters, M.S. and Olivier Bodenreider, M.D., PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
{lee.peters|olivier.bodenreider}@nih.gov

Motivation
The National Drug Code (NDC) is a universal product identifier for human drugs in the United States. NDCs can be found on drug packaging and are listed on the Food and Drug Administration (FDA) Structured Product Labels (SPLs). NDCs are used by pharmacists for inventory management and for drug information exchange, as well as in claims data. From a clinical perspective, however, NDCs are too detailed (e.g., packaging information) and need to be mapped to clinical drugs or ingredients (e.g., for medication reconciliation). It is therefore not surprising that the most used query from users in the RxNorm Application Program Interface (API) [https://rxnav.nlm.nih.gov/RxNormAPis.html] is to map NDCs to the corresponding RxNorm concepts. RxNorm currently contains over 200,000 NDCs, most of which come from SPLs and the rest from First Data Bank. However, until recently, the RxNorm API did not expose the properties associated with NDC (extracted from SPLs). The new API function described here now provides this information.

NDC Properties
NDCs are multi-segment identifiers, with codes for manufacturer, drug, and packaging information. NDCs come in multiple flavors, namely 3-segment, 2-segment (manufacturer-drug), and an 11-digit format combining the 3 segments. NDCs are associated with other codes, such as the Abbreviated New Drug Application (ANDA) code and the identifier for the SPL (“SPL set id”). Other information linked to the NDC include pill characteristics (e.g., color, shape, size), as well administrative information (e.g., marketing category, labeler, schedule). The RxNorm API function getNDCProperties takes as input an NDC (any flavor), an SPL set id or an RxNorm identifier and returns the properties associated with the NDC (for all NDCs from SPLs). While most of the properties are extracted from RxNorm, packaging information is acquired from DailyMed. NDC properties for 0781-1506-10 (Atenolol 50 MG Oral Tablet) are listed in Table 1.

Table 1 – NDC properties returned by the RxNorm API

<table>
<thead>
<tr>
<th>Property</th>
<th>Example</th>
<th>Property</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>RxCUI</td>
<td>197381</td>
<td>Labeler</td>
<td>Sandoz Inc</td>
</tr>
<tr>
<td>2-segment identifier</td>
<td>0781-1506</td>
<td>Packaging</td>
<td>1000 in 1 BOTTLE</td>
</tr>
<tr>
<td>3-segment identifier</td>
<td>0781-1506-10</td>
<td>Shape</td>
<td>ROUND</td>
</tr>
<tr>
<td>11-digit NDC</td>
<td>00781150610</td>
<td>Color</td>
<td>WHITE</td>
</tr>
<tr>
<td>ANDA</td>
<td>ANDA073025</td>
<td>Imprint</td>
<td>GG263</td>
</tr>
<tr>
<td>SPL set identifier</td>
<td>b90115a-faac-4244-94bc-c1ef2f88aa38</td>
<td>Size</td>
<td>7 mm</td>
</tr>
<tr>
<td>Marketing Category</td>
<td>ANDA</td>
<td>Score</td>
<td>2</td>
</tr>
<tr>
<td>Marketing Status</td>
<td>active</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing Effective Time</td>
<td>19910917</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label Type</td>
<td>HUMAN PRESCRIPTION DRUG LABEL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use Cases
The RxNorm API function getNDCProperties was designed to support several use cases.
- Finding the NDC properties for a clinical or branded drug. For a clinical or branded drug in RxNorm, the NDC properties of all the NDCs for that drug can be returned from a single call to the API.
- Finding the NDCs and properties for a Structured Product Label. The API function will return all the NDCs and their properties for a given Structured Product Label.
- Finding the characteristics of oral solid dose form medications (pills). The NDC properties returned include color, shape, size, and imprint code.
- Finding the labeler/manufacturer information for a drug. For a given NDC, the labeler/manufacturer is returned. By specifying an RxNorm identifier as input, the labeler/manufacturer information is returned for each associated NDC.

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.
Use of Freetext Clinical Reports for Prediction of 30-Day Psychiatric Readmissions

Jose Posada, MS1,3, Lingyun Shi, MS1, Ye Ye, MS1, Neal Ryan, MD2, Frank Ghinassi, MD PhD2, PhD, Henk Harkema, PhD4, Fuchiang (Rich) Tsui, PhD1
1Department of Biomedical Informatics and 2Department of Psychiatry University of Pittsburgh, Pittsburgh, PA; 3Electronics and Telecommunication Engineer Program Universidad Autonoma del Caribe, Barranquilla, Colombia; 4Nuance Communications®, Pittsburgh, PA)

Abstract: We assessed the contribution of free-text clinical reports in predicting 30-day psychiatric readmissions. We extracted 5,495 clinical findings from 491 psychiatric visits during one month using cTAKES and identified top predictive terms using feature selection. We built a Bayesian classifier using the leave-one-out cross-fold validation (LOOCV) approach. The model had an area under receiver operating characteristic curve (AUROC) of 0.75 (95% CI 0.66-0.84). Our findings indicate that clinical reports appear to be valuable in predicting 30-day psychiatric readmissions.

Introduction: According to the substance abuse and mental health services administration (SAMHSA) 43.1 million of adults (18.1%) in the US have experienced some form of mental illness1. Hospitalizations for psychiatric illnesses are increasing at a faster rate than any other type of hospitalization1. For hospitalized psychiatric patients, 15% of all discharged patients are readmitted within 30 days and these readmissions are associated with higher costs than any other readmission cause1. Moreover, CMS releases individual hospital readmission rates to the public as an indicator of quality of care. One of the key steps to reduce 30-day readmission rates is to identify patients at high risk of 30-day readmission; however, existing studies2 have used only structured psychiatric data (e.g., demographics) and have achieved AUROCs less than 0.63. In this study, we hypothesize that unstructured clinical reports can be utilized to better predict 30-day psychiatric readmissions.

Methods: We retrieved all the clinical reports from one month of inpatient visits at the Western Psychiatric Institute and Clinic of the University of Pittsburgh Medical Center (UPMC) during December 2007. We used the visits from the following month (January 2008) to determine readmission cases. We utilized cTAKES3, an open source natural language processing tool, to extract signs and symptoms from the clinical reports and encode them into one of three values: Present, Absent and Missing. To predict readmission visits, we employed a naïve Bayesian (NB) classifier because it can handle uncertainty and has performed well in our prior work4. We used LOOCV to evaluate the classifier performance. We chose the top 50 features automatically ranked by information gain to train the classifier, which was based on the balance between model complexity and available sample size.

Results: The pilot dataset comprised 941 clinical reports from 491 inpatient visits (36 readmission cases and 455 controls; 7.3% readmission rate). cTAKES extracted 5,495 unique clinical findings from the reports. The dataset comprised three report types: discharge summaries (47.5%), history and physical (52.2%), and progress notes (0.3%). The final classifier had an AUROC of 0.75 (95% CI 0.66-0.84). Top-performing features included feeling suicidal, hallucinations and group psychotherapy. These features were also identified as key factors for readmissions in a prior study by Machado5.

Conclusion: In this pilot study, we found that free-text clinical reports appear to be valuable in predicting 30-day psychiatric readmissions with AUROC 0.75. The limitation of the study is the small number of readmission cases. We plan to collect more cases, refine cTAKES, and retest our hypothesis in a larger scale study.

References

Systemic identification of analytes tracked as outcomes in RPM clinical trials

Gouri Prakash, BE, MISM, PGDCR

1 CitiusTech Inc., USA

Abstract

Remote Patient Monitoring (RPM) technology facilitates delivery of care outside conventional clinical settings, such as the patient’s home. Waived lab tests provide opportunity to conduct a limited amount of self-monitoring. This poster evaluates the extent to which RPM clinical trials measure values of analytes designated as waived.

Introduction

RPM technologies support or enable care models that reduce avoidable re-admissions, promote transition of patient care from the hospital to home and empower patients with chronic conditions to regularly monitor the severity of their illness. RPM clinical studies evaluate the efficacy with which RPM technology improves clinical and financial outcomes. In order to assess clinical outcomes, some of these trials measure values of one or more analytes. This study was undertaken to systemically identify the analytes measured during the conduct of RPM clinical studies and to systemically identify which of these analytes were also on the list of waived analytes. Waived analytes by definition can be measured by lab tests that are approved by the FDA for home-based use. The objective of the study was to gain insight into the utilization of analytes on the waived list in RPM clinical studies.

Methods

The keyword “Remote Patient Monitoring” was used to retrieve studies from the ClinicalTrials.gov portal. The result set comprised of 238 clinical studies. The information provided in the outcomes field for each study was parsed to isolate analytes that were measured during the conduct of trials. The Component field in the LOINC dataset was used for the detection of analytes embedded within the outcomes section for each study. This list of analytes was parsed further to identify whether any of these analytes were designated as waived which was done by leveraging the he list of waived analytes as a reference to arrive at the matches between the two lists.

Results and Discussion

Out of 238 clinical trials, there were 27 trials that measured analytes designated as “waived”. 20 of these 27 trials (74%) evaluated the efficacy of RPM in improving outcomes for diabetic patients (Type 1 diabetes, Type 2 diabetes or gestational diabetes). The analytes measured to evaluate these outcomes were glucose, HbA1c, lipid profile, inclusive of cholesterol and triglycerides, and in one case fructosamine - all of which have been designated as waived by the FDA. Out of 123 waived analytes, there were 13 analytes (10.56%) that were measured to evaluate clinical outcomes in RPM clinical studies. From an informatics standpoint, as the outcomes of RPM clinical studies was provided in natural language, the LOINC dataset and the waived analyte list served as controlled vocabularies, in the detection of waived analytes embedded within the outcomes of interest in RPM clinical studies.

From a clinical standpoint, since waived lab tests by definition have regulatory approval for home-based use, these tests can potentially be leveraged in the model of care enabled by RPM. Match on just 13 analytes indicates under-utilization of waived lab tests within the model of care enabled by RPM. More than two-thirds of the analytes on the waived list comprise of biomarkers of normal biological processes, which implies that these waived tests can be used for monitoring purposes. Alternatively, a hybrid model of care which makes use of tenets of RPM and is enabled by the availability of waived tests can be hypothesized. For example, although none of the RPM clinical studies measured outcomes for patients suffering from gout, care providers can put such patients on care plans designed to self-monitor the levels of uric acid, a waived analyte. Waived lab tests can potentially scale the model of care enabled by RPM for newer target patient populations and improve the efficacy of the existing model of care.

References

Usability of a Provider-Facing Breast Cancer Prevention Toolbox

Jordan L. Price, BS¹, William Ueng, BS¹, Austin M. Coe, BA¹, Raven David, BA², Meghna Trivedi, MD², Katherine D. Crew, MD¹,²,³, MS, Rita Kukafka, DrPH, MA¹,²

¹Mailman School of Public Health, Columbia University, New York, NY; ²College of Physicians and Surgeons, Columbia University, New York, NY; ³Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY

We developed a provider-facing toolbox for breast cancer risk assessment and prevention in the primary care setting. The toolbox informs providers about the risks and benefits of genetic testing and chemoprevention and facilitates patient engagement. We conducted usability testing to identify themes related to ease of use, content, and navigation. Further evaluation in a controlled trial will determine if the toolbox improves the identification and referral of women at increased risk for breast cancer.

Introduction

Breast cancer poses a public health burden as the most commonly diagnosed and second leading cause of cancer-related deaths among women in the United States. Several known risk factors affect breast cancer development, and primary care providers (PCPs) may employ risk assessment tools to calculate patient risk. However, available tools and prevention methods are not routinely employed in practice. This study aims to assess the usability of a provider-facing Breast Cancer Prevention Toolbox targeted for providers in the primary care setting.

Methods

The toolbox provides modules on chemoprevention and genetic testing consisting of key concepts, guidelines, and multimedia case-based learning. The “My Patients” section displays relevant information for patients at high breast cancer risk. The patient’s 5-year risk according to the Gail model is shown with a link to a patient-generated action plan with family history, patient-identified pros and cons for risk-appropriate preventive services, and action items. The toolbox can be accessed through the electronic health record or directly on the web. Seven PCPs were identified through Columbia University Medical Center (CUMC) in New York, NY. Testing sessions were recorded and lasted approximately one hour during which subjects were asked to “think aloud” while navigating the toolbox. A monitor facilitated each study while an observer documented comments and reactions. Qualitative data analysis promoted the identification of themes related to the program’s Ease of Use, Content, and Navigation.

Results

The identified issues in each of the thematic areas are summarized in the table (Table 1) below.

Table 1. Themes identified and addressed during the usability studies specific to the toolbox.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Location</th>
<th>Problem Description</th>
<th>Resulting Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of Use</td>
<td>BNAV Dashboard</td>
<td>Unclear about what can be found by accessing each link</td>
<td>Hover over a link and a drop down menu appears with site subsections</td>
</tr>
<tr>
<td></td>
<td>My Patients</td>
<td>Links for patient risk and action plan often overlooked/not obvious</td>
<td>PDF icon in place for each patient risk and action plan link</td>
</tr>
<tr>
<td>Content</td>
<td>My Patients</td>
<td>Unfamiliar with the average risk score relative to general population and recommended steps in referral process</td>
<td>The average patient risk score and referral recommendations added to My Patients module</td>
</tr>
<tr>
<td>Navigation</td>
<td>BNAV Dashboard</td>
<td>My Patients module link should be distinguished from education links</td>
<td>My Patients link and icon placed prominently in homepage, distinct from education modules</td>
</tr>
</tbody>
</table>

Conclusion

The toolbox addresses a gap in PCP breast cancer risk assessment and prevention management. The themes we identified provide insight to achieve optimal user engagement. Further evaluation in a controlled trial will determine if the toolbox improves the identification and referral of women at increased risk for breast cancer.
A platform for rapid data sharing and visualization to contribute towards future epidemic preparedness within communities in Sierra Leone

Nuri Purswani, PhD, Meenal Pore, PhD, Purity Mugambi, BSc, Samuel Karumba, BSc, Kala Fleming, PhD, Reginald Bryant, PhD

1IBM Research Africa, Catholic University of Eastern Africa, Nairobi, Kenya

*Equal contributors to this work

Introduction

A vital lesson learned from the 2014 Ebola outbreak in West Africa is the need to develop tools that enable government agencies and NGOs to rapidly share data in the event of an emergency. In Sierra Leone, the majority of data in health facilities is collected through paper forms: a time-consuming process that often leads to information of dubious quality. In addition, health facilities often do not report essential information, which hinders the process of data-driven decision-making in policy. The gradual transition towards mobile-based data collection platforms provides a promising opportunity for improving data quality. Furthermore, datasets from different organizations (e.g., NGOs and Government) need to be interoperable so as to enable end users to quickly obtain insights from different reports without the need for expert knowledge or long delays manipulating files before information is available for action.

Architecture

To address the challenges above, we have developed a web-based platform enabling the following key functionalities: ingestion of tabular data, unstructured text management (mobile) and real-time API integration with mobile collection platforms like Open Data Kit, Commcare and Kobo Toolbox; geotagging, deduplication and matching of related fields across multiple datasets (to match multiple naming conventions to the same location); visualization dashboards for uploaded datasets and in-built analytics to enable data exploration; alerts dashboard to flag areas in risk based on user defined indicators (i.e. the user selects indicators and the system triggers alarms based on pre-defined conditions) and natural language processing tools to generate word clouds and provide rapid situational awareness in different parts of the country.

We developed back-end analytics to flag important pieces of information for epidemic preparedness in Sierra Leone. The analytics are broken down into: (1) A trigger system flagging alerts at health facilities when diseases such as measles, yellow fever, typhoid or maternal deaths are reported; (2) An alarm system for infection detection and prevention and also for the detection of events that relate to un-sufficient or deficient control structures, inventories or practices; (3) An alarm system detecting section death alerts that exceed predetermined threshold (e.g., 17 deaths per 1000 population per month). In addition, the developed analytics also flag health facilities that are not reporting data consistently. Finally, to verify reports from health facilities, we developed a dashboard that enables officers to produce real-time reports automatically geotagged through the KoBo Toolbox app. The real-time information input by users is visualized through a second dashboard that contains a word cloud and highlights key topics reported by date and geography.

Pilot

The pilot for the tool is currently taking place in Sierra Leone. We will present the results during the workshop.

References


S2CR3UM: A Solution to the In Silico Relevance, Reliability & Reproducibility Conundrum

Sarah B. Putney, JD, MA, Andrew White, PhD, Janos Hajagos, PhD, Joel H. Saltz, MD, PhD, Jonas Almeida, PhD, Mary. M. Saltz, MD; Department of Biomedical Informatics, Stony Brook School of Medicine, Stony Brook, NY

Introduction
Informaticists face daunting challenges with data management. Stony Brook Department of Biomedical Informatics created a quality control program to improve reliability, reproducibility, and relevance of data products.

Methods and Objectives:
We defined two synergistic goals for this project: a) to develop, implement, and iterate on a Toolkit for Quality improvement and ongoing quality control (QC) for data analytics; and b) to transform the culture to support a nimble and productive team. The program meta-process is organized around defined roles and responsibilities, and coordinated across time and technical systems, to deliver data products meeting scientific standards of relevance, reliability, and reproducibility. People & Communications: Roles and responsibilities were defined for the team and all developed tools and shared ideas for process improvements, based in part on The Checklist Manifesto. Building on this foundation, the group drafted of a core set of checklists to guide workflow and to promote accuracy and consistency. Borrowing from the model of agile software development, a rapid cycle of work-review-correct/revise, a set of “scrum”s (informal but focused meetings) began: thrice-weekly scrums where developers report on current work; more inclusive weekly scrums at which project priorities are set and data products in process are critiqued; and a weekly “super” scrum for program leaders. Program leaders reinforce expectations for attendance and participation. Systems for Data, Code & Data Products: A shared infrastructure was adopted to promote interoperability, retrieval, version control, reproducibility, etc. OpenProject (https://www.openproject.org/) is used for project management with projects and tasks defined using a simple structured format (Problem, Materials, Methods, Observations, Conclusions, Application.). Github (https://github.com/) is the designated repository to store software code for work in process and finalized. Analogously, BOX (https://www.box.com/) serves as the repository of data products at all stages of development, made searchable through a naming convention that included task #, completeness level, a 4-5 word description of the product, and file format. Links to storage of final products and code are placed in the OpenProject tracker. A 5 level grading system, with a checklist for each completeness level, was developed. The final and most complete level (5), includes criteria such as references to data provenance, a cross check for data veracity, the number of populations and subgroups analyzed, and geographic or service provider scope. Interactive tools for data visualization and exploration, such as integration of patient addresses to Google Street View were developed to help guide understanding of results. Standard templates are used to promote brand identity and consistency. Monthly project review meetings are conducted, engaging subject matter experts as needed, to ensure relevance and completeness of data products. Observations: Nine months after implementation review shows new data products to be easily retrieved, consistent, reproducible and well branded. This effort encompassed dataset testing and validation to ensure accuracy of combined in patient and outpatient data including laboratory test, location and medication information. Insights gained in this dataset quality process allowed us to collaborate with Cerner to optimize usefulness of the information obtained from our newly deployed Healthe EDW data warehouse and to target a variety of quality efforts including infection control, systematic targeting of Clostridium difficile (C. diff) testing along with analysis of Suffolk County New York population patient health quality indicator and readmissions related metrics. Challenges include consistent adhere to standards and checklists, but changes in workflow have made consistent QI a vital part of departmental culture.

Conclusion
We conclude that by changing culture around data quality and by using a Toolkit for Quality, scientific integrity and in silico quality control can converge in a manageable, affordable, and productive workflow in an academic setting.

References
Topic Visualization of Online Cancer Discussion Forums
Jin Qu, BS, Annie T. Chen, PhD
Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, Washington, USA

Abstract
Online cancer discussion forums are a rich source of health information and can potentially have a great impact upon people’s medical decision-making. However, given the dynamic nature of online discussion data, it is a challenge to extract valuable insights from these data. We present a prototype visualization to aid exploration of people’s online information and support exchanges.

Introduction
Recent literature has observed that social media and online communities are reshaping the healthcare landscape. As people’s interest in seeking and exchanging online health information increase, it is of great importance to develop a high-level understanding of this information, including topics that draw the most attention, and differences among online communities. We have leveraged text mining and visualization techniques to prototype a visualization tool that facilitates exploration of topics in online cancer-related discussions.

Methods
Data Collection: We used Wget to extract webpages from CancerCompass, a website featuring cancer-related message boards. Our corpus is comprised of 152,378 posts (from July, 2001 to February, 2016), covering a wide range of topics including cancer diagnosis, treatment and recovery.

Data Processing and Topic Mining: Discussion content was extracted from the raw HTML files and pre-processed using BeautifulSoup. We then utilized MALLET to perform topic mining using Latent Dirichlet Allocation (LDA). In addition to the default stop word list, the top 100 highest frequency words were removed from the corpus to increase topic differentiation. This list is available upon request.

Data Visualization: The visualization was developed using d3.js. We set three design goals:
1. To visually represent topic prevalence. The topics are represented using circles, where the radius of each circle is proportional to the proportion of each topic within the corpus. The use of circle size to indicate prevalence has been used in previous visualizations of textual data including discussion forums and newsgroups.
2. To facilitate association of topics by semantic category. Different colors are used to represent topic categories: e.g. body, habits, information, and medical events.
3. To provide detailed summary statistics and links to posts associated with each topic. When a user hovers over a circle, summary information and a hyperlink list to posts will be displayed (in progress).

Preliminary Results
A visualization of the Head and Neck Cancers message board is displayed in Figure 1. Discussion about feeling is the most prevalent, followed by spirituality, doctors, lymph and treatment effects.

Conclusion
We have demonstrated a visualization tool to present text mining results. This tool can be used by researchers and health consumers to explore the dynamics of people’s online health information and support exchange behaviors. This tool can also be used by patients, caregivers, and biomedical informaticians in shared decision-making contexts.

References
Mobile Maternal Health Applications in Developing Countries

Yuri Quintana¹, Ph.D., Jennifer McWirther², Ph.D., Melek Somai¹, MD MPH, Michelle Hacker¹, ScD, MSPH, James Gray¹, MD,
¹Beth Israel Deaconess Medical Center, Boston, MA, USA
²University of Guelph, Guelph, Ontario, Canada

Introduction

A long-term goal for maternal and newborn health is to prevent and reduce newborn deaths and morbidities. In 2013, 4.6 million children died (74% of all under-5 deaths) within the first year of life¹. An estimated² 270,000 newborns die during the first 28 days of life every year from congenital anomalies. To address this global health problem, the United Nations (http://www.un.org) included maternal health in their Millennium goals and the new Sustainable Development Goals. Governments, non-profit groups and research groups worldwide have launched initiatives to improve outcomes. Mobile technologies hold the potential to improve maternal and newborn health by supporting access to education, improving coordinated care, and facilitating patient support. Some of these systems have been designed to improve maternal outcomes such as increasing patient visits for antenatal care, improving healthier behavior in patients such as avoiding smoking and alcohol, and improving vaccination rates of newborns.

Method

We searched the literature for any intervention delivered during pregnancy or up the first 28 days, where the intervention was delivered digitally (not voice alone) via mobile phone or smart phone. The effect of the mobile intervention had to be discernable from other intervention components. Computerized literature searches were performed with no date or language delimiter using the following literature databases: PubMed/MEDLINE (NCBI), Embase (Elsevier), Cumulative Index to Nursing and Allied Health Literature (EBSCO), Cochrane Central Register of Controlled Clinical Trials (EBSCO), Web of Science (Thomson), Popline (K4Health). The search was focused on mobile applications on mobile devices and platforms for maternal health. The search was constrained to the title and abstract words and controlled vocabulary terms, when available, and to peer-reviewed journal papers. One author completed the electronic search. All items in the final list were reviewed by two or more authors.

Observations and Discussion

We found 1,680 references using the search strategy of which 32 met our inclusion criteria. Of these, the studies used the mobile phone applications for: reporting test results from clinics to patients via mobile apps (n=1), patient reminders for upcoming clinic visits via mobile alerts (n=5), patient-reported medical status (n=5), patient education via mobiles for behavior change (n=21). The sample size of participants in the studies ranged from 68 to 2550. There were mixed findings for the utility of mobile health applications in the context of maternal and child health. A wide range of hardware- and software-based technologies have been developed to address various aspects of maternal and newborn health. Despite the many programs and publications, very few studies took the form of randomized control trials. Implementation problems included, for example: lack of availability of mobile phones by patients in some developing countries; issues around staff training and incentives; and a lack of health literacy measurements for educational content. Studies, and the applications considered within, were also largely uninformed by health behavior models or theories. Evaluation problems included not having a large enough sample size for statistical significance, a high number of dropouts and the lack of qualitative measures or validated assessment tools.

Conclusions

A variety of economic, geopolitical, sociocultural, and technical factors need to be addressed to achieve successful deployments of mobile maternal health applications and platforms. Future studies seeking to evaluate such tools should use randomized control trials in order to understand effects. The success of these systems is very much dependent on the design and implementation that is appropriate for local contexts.

References

Serious digital game for heart failure self-management by older adults: A pilot study

Kavita Radhakrishnan, RN PhD MSEE, Paul Toprac, PhD, Matt O'Hair, MA, Paul Bradley RN MSN, Miyong Kim RN PhD FAAN, Mike Mackert, PhD

1The University of Texas, Austin, TX, USA; 2The University of Texas Game and Multimedia Applications (GAMMA), Austin, TX, US

Introduction

Effective self-management of heart failure (HF) which has an estimated worldwide prevalence of over 23 to 41 million can lower mortality and hospitalization rates. Serious games as a modality is yet to be explored for disease management among older adults. In this abstract, we present the development and testing of a serious game prototype by an inter-disciplinary team from nursing, computer game programming, usability science and communication for improving HF self-management knowledge and efficacy among older adults.

Method

The game genre of casino slots was selected based on a survey of 34 community based HF nurses on their and their patients’ preferences for playing digital games. HF education content adapted from the Heart Failure Society of America (HFSA) evidence-based guidelines and tailored to low literacy levels was placed strategically throughout the game using game-based instruction strategies. Participants 55 years or older and admitted with a HF diagnosis were recruited from three out-patient HF clinics based in central Texas, USA. Usability of the digital game was tested on seven HF individuals through guided observations by a trained Research Assistant, and a game usability questionnaire influenced by the Serious Game Usability Evaluation and the interest / enjoyment subscale of the Intrinsic Motivation Inventory (IMI) questionnaires. During the beta-testing of the game with 19 participants over four weeks, a pretest-posttest design was used to test improvement in HF knowledge and self-management which was measured using validated instruments of the Atlanta Heart Failure knowledge Test (AHFKT) and Self Care Heart Failure Index (SCHFI). A post-game survey was used to assess participant perceptions on the game. The study received institutional review board approval from the University of Texas – Austin.

Results and Evaluation

Of the seven community-residing older adults (57% females) with HF who participated in the usability testing of the digital game, 100% found the game easy to play, 100% found it enjoyable and 86% found it helpful to learn about HF. Of the 19 participants in the beta-testing of the game (90% were males, 84% were Caucasian, 72% lacked a degree, 58% older than 70 years age, 58% diagnosed with heart failure for more than 10 years), 90% found the game interesting, enjoyable and easy to play. Participants with lower education level preferred games to any other medium for receiving information (p= 0.02). Playing our game resulted in significant improvement in HF knowledge (p=0.007), non-significant improvement in HF self-maintenance (p=0.11) and no difference in HF self-confidence scores.

Discussion

Developing a serious game that is satisfying and acceptable to older adults with HF is feasible. Moreover, game-based instruction was established as an effective manner to impart knowledge on chronic diseases such as heart failure, especially for participants with lower education levels. The next steps include studies to explore the impact of gamification principles for HF self-management behavior modification and improving older adults’ heart failure outcomes, including self-management, re-hospitalizations, and quality of life.

References

Parsing complex microbiology data for secondary use

Protiva Rahman, BS¹, Albert M. Lai, PhD¹, Courtney L. Hebert, MD, MS²,³

¹Department of Computer Science and Engineering, ²Department of Biomedical Informatics, ³Department of Internal Medicine, The Ohio State University, Columbus, OH

Introduction

Clinical microbiology culture and antibiotic susceptibility data are a rich source of information required for a variety of secondary use applications such as epidemiologic studies of antimicrobial resistance. Our study team is working on using clinical microbiology data to develop antibiotic prescribing clinical decision support.¹ Unfortunately, culture and antibiotic susceptibility results are often reported in complex, semi structured or free text form, making it difficult to use them for analysis. Cultures from abdominal-biliary (AB) sources are especially complex because there are multiple types of cultures results (e.g. anaerobic, fungal) and a wide variety of organisms. Prior work in this area has focused on rule-based approaches, natural language processing, and often on blood cultures.²,³ The objective of this project was to parse and annotate each unique organism name from a free text field of AB cultures and identify any antibiotic susceptibility results or evidence of resistance mechanisms (e.g. presence of penicillinase production) that occur in this free text field. By annotating the organisms with a unique concept identifier we are able to add additional, useful information to the dataset, such as Gram staining, type of growth patterns, and family.

Methods

Adult patients admitted to an inpatient unit with a diagnosis of AB infection, with a positive culture from a relevant site in the first 4 days of their hospitalization to The Ohio State Wexner Medical Center from 1/1/2009 to 1/1/2014 were included for a total of 625 patients and 1023 unique cultures. The UMLS metathesaurus⁴ was used to identify organism name. Our method included: 1) Tokenizing the result text and removing. 2) Using regular expressions to extract colony count and other culture data (e.g., resistance mechanisms). 3) If a token did not match any predefined regular expression, we checked to see if it matched the concept type of ‘bacteria, virus or yeast’ in the UMLS. If it did we continued consuming tokens to find the longest matching string that matched the semantic type identifier for a pathogen. 4) In the case where an organism was not fully speciated (e.g. Escherchia/Citrobacter), we assigned the code for the common parent. 5) If the token matched the UMLS semantic type of “Antibiotic,” we look at the surrounding tokens with a span of two to see if they match a set of positive words: [e.g., confirmed, produces, covers] or a set of negative words: [e.g., negative, resistant] and accordingly assigned true or false. We developed our algorithm on 300 cultures and tested it on the remaining 723. Our subject matter expert (CH) manually validated a set of 200 parsed results, for which we report the precision (true positive/(true positive + false positive)), recall (true positive/(true positive+false positive)) and F1 score (2*precision*recall / (precision + recall)).

Results

For identifying the correct organism precision was 0.91, recall 0.95, and F1 0.93. Penicillinase production was identified and related to the correct organism with a precision of 1, recall 0.95, and F1 0.97. Free text antibiotic susceptibilities were identified and related to the correct organism with a precision of 0.92, recall 0.93, and F1 of 0.93. The major challenges we faced were in dealing with not fully speciated organisms and relating parsed data to the correct organism in a poly-microbial infection.

Conclusion

We achieved high F1 scores using our technique of longest matching and regular expressions customized to our data. While some techniques are specific to our data, the ideas can be replicated for other free text microbiology data. This kind of work is vital in order to use these complex data for secondary use applications.

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the NIH under Award Number R01AI116975

References

**Rural Health Informatics: Using EHR Data for Systems Based Improvements in the Emergency Department**  
**Gilbert H Ramirez, MD., James P Killeen, MD.**  
**University of California, San Diego**

**Introduction:** El Centro Regional Medical Center is a safety net hospital with an annual census of 52,000. In 2015, one out of every 25 patients left the emergency department without evaluation by a physician. This is double the rate of the previous year. This represents a symptom of increased waiting times, patient volume and is a risk to hospital safety.

**Methods:** Using hospital EHR data and physician billing data we plotted arrival and screening curves for the patient census. These curves were compared with time to bed and average expected physician workflows. We adjusted this data to model intake workflows as patients per hour. We assumed an inefficient system when curves did not intersect. From the combination of these curves, assumptions of physician workflow, fatigue and department triage efficiency were made. This data was used to make staffing and workflow adjustments to the intake process. Over the next few months we will observe the difference in left without being seen rates after implementation.

**Results:** From the model, we were able to identify areas that required systems based process improvements. During the morning hours a large patient surge quickly overwhelms the intake process of the emergency department. In addition, physicians are unable to meet screening demands despite working beyond their expected capacity. From these observations, we implemented a new triage and provider staffing model. After a two-month adjustment period the left without being seen rates will be compared. We hypothesize that the model is useful as an easy data-driven tool to identify targets to reduce left without being seen rates in the emergency department.

![Figure 1. Identifying need for provider staffing.](image)

**Conclusion:** Hospitals with access to limited data streams can still identify and target process improvements. These simple models can potentially avert risk and improve patient care by re-allocating limited resources to periods of need. Ways to quickly and transparently identify these targets may improve access to patient care in these communities.
Patient Identity Matching for Health Information Exchange between Poison Control Centers and Emergency Departments

Pallavi Ranade-Kharkar, PhDc, MS1,2, Darren Mann, BS1, Heather Bennett MPA, Barbara Crouch, PharmD, MSPH2, Guilherme Del Fiol, MS, PhD2, Sidney N. Thornton, PhD1,2, Mollie Cummins, PhD, RN, FAAN2
1Intermountain Healthcare, Murray, UT; 2University of Utah, Salt Lake City, UT;

Introduction
Poison control centers (PCCs) routinely collaborate with emergency departments (EDs) to provide care for patients exposed to poisoning. However, PCCs and EDs use disparate information systems, communicate using a telephone-based process and do not routinely share data electronically1. Electronic health information exchange (HIE) has the potential to improve the current process by enabling appropriate and timely routing of information, shared documentation for patients, and addressing ambiguity through exchange of standardized documents. Patient identity matching is both a crucial first step and a known barrier in HIE2. Automated patient identity matching based on demographic data collected by PCCs is the optimal workflow. However, the completeness and validity of PCC data for the purpose of patient identity matching is unknown. The objective of this project is to determine the feasibility of automated patient identity matching between the Utah Poison Control Center (UPCC) and Intermountain Healthcare EDs.

Method
We performed Cross-Community Patient Discovery (XCPD) transactions for successful PCC referrals to Intermountain Healthcare EDs across the State of Utah, that occurred in the year 2015 (n=1154), in order to determine the extent of successful patient identity matching between UPCC and the EMR at the EDs. In a successful PCC referral to an ED, the PCC recommends that a patient go to the ED for further medical evaluation and treatment, and the patient complies with that recommendation. Therefore, all cases in the sample involved patients known to Intermountain Healthcare. We calculated the number and percentage of cases with a successful XCPD match, and conducted a manual review of the dataset with unsuccessful match, to determine the cause.

Results and Discussion
The patient attributes currently captured in ToxiCALL® result in only 7% of the patients matching between the UPCC and the EMR at the EDs. Manual review of the dataset revealed that the addresses included the city, state and zip code and did not include the street address for 100% of the cases. Patient date of birth (DOB) was captured for 305 (26.64%) records only. The phone number was captured for the caller and not the patient for the entire dataset. While it is likely that the phone number of the caller matches that of the patient in a subset of the cases, it is unlikely that it matches for all cases. All of the matched records included the patient DOB. We concluded that the primary reasons for a low match rate included: 1) missing DOB; 2) incomplete address; and 3) phone number captured for the caller (instead of the patient). Going forward, the UPCC staff will be trained in an enhanced workflow where the distinguishable characteristics for a patient match, namely, patient DOB, address and phone number are captured consistently, in a standard format. We will re-assess the extent of successful patient identity matching after workflow modification, and a deferred patient linkage process will be used to resolve patient identities for the exceptional cases in which automated patient identity matching is not successful.

Conclusion
Data quality of patient demographics at the intake of poisoning cases is critical to patient identity matching between PCCs and EMRs at the EDs. Low patient identity match rates can be remedied by improved workflow and training to capture the distinguishable characteristics for a patient match. Analyzing HIE barriers and facilitators at design time while connecting heterogeneous systems not designed for HIE, can improve feasibility and efficiency of HIE.

References
Detecting Anomalies in Alert Firing in Clinical Decision Support (CDS) Systems using Anomaly/Outlier Detection Techniques

Soumi Ray, PhD¹, Dustin McEvoy, BS², Jan Marie Andersen, MS², Adam Wright, PhD²
¹Northeastern University, Boston, MA; ²Brigham and Women’s Hospital, Boston, MA

Introduction
Anomaly detection techniques are commonly used in many different domains such as credit card fraud, data/network security, and meteorology. CDS systems which facilitate clinicians in their decision making process suffer from malfunctioning alerts which often go undetected for months or even years before being recognized. We propose the application of two algorithms for detection of anomalous triggering of alerts within CDS.

Method
Anomalies can be generally classified into two classes: point and breakout anomaly. Point anomaly refers to a sudden spike or drop in time series. Breakout is characterized by two steady states and an intermediate transition period. Point anomalies can be detected using the Seasonal Hybrid Extreme Studentized Deviate algorithm (S-H-ESD)¹. This algorithm approximates the underlying trend in a long-term time series using a piecewise median method. Breakout anomalies can be detected using a novel statistical technique, called E-Divisive with Median (EDM)². Alerts originating from Brigham and Women’s electronic health records are used in this study. Point anomaly detection was applied to alerts originating from patients with coronary artery disease who were not on aspirin (a deviation from US Preventive Services Task Force recommendation). Breakout detection was applied to alerts originating from immunocompromised adults <65 years old who did not receive Pneumococcal Conjugate Vaccine (PCV) according to existing guidelines.

Result
Figure1(a) shows the point anomalies detected by S-H-ESD, which start around 23rd May 2012. These anomalies were identified and validated as a malfunction in the drug classification service, which caused the rule to fire for all patients with CAD even if they were already taking Aspirin. In Figure1(b) the red dotted lines denote the start time of breakouts detected by the EDM and the ordinates correspond to the approximate mean of the previous window. We compute the pairwise p-value using Wilcoxon test shown in blue above the red dotted line, which gives an estimate of the magnitude of the shift at that corresponding timestamp based on data in the time intervals (determined by start times of breakouts) immediately preceding and succeeding that breakout time. During the first mean-shift around 22nd July 2015, a change was noticed in PCV alert firing. Changes in the rule designed to fire PCV alert even for patients already vaccinated, could potentially explain this breakout.

Conclusion
We used two novel anomaly detection algorithms on alert firing data and preliminary results show that they are able to detect the approximate dates of anomalous alert triggers which correspond to known CDS system malfunctions. These algorithms could aid early detection, but further testing on other CDS alerts is needed for validation.

References
Utilization of Cost Effective Tools for Queries on Healthcare System Stress

Taylor Read MS\textsuperscript{1}, Elizabeth White\textsuperscript{1}, Neelima Karipineni MD, MMSc\textsuperscript{1,2,3}, Maksim Ignatov\textsuperscript{1}, Mahesh Shanmugam\textsuperscript{1}, Beatriz Rocha MD, PhD\textsuperscript{1,2,3}, J Perren Cobb MD\textsuperscript{4}, Saverio Maviglia MD, MS\textsuperscript{1,2,3}, Roberto Rocha MD, PhD\textsuperscript{1,2,3}, Sarah Collins RN, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Brigham and Women’s Hospital, Boston, MA; \textsuperscript{2}Harvard Medical School, Boston, MA; \textsuperscript{3}Partners Healthcare System, Wellesley, MA; \textsuperscript{4}University of Southern California, Los Angeles, CA

Introduction
National preparedness efforts for public health emergencies are important in order to address challenges to healthcare and communications systems. The United States Critical Illness and Injury Trials Program for Emergency Preparedness, in partnership with the Office of the Assistant Secretary for Preparedness and Response, aims to increase the capacity for multi-site, prospective data collection during emergent events. Success is dependent on the management of large participant lists using widely-available and open-source software including REDCap and an analytics platform. We implement these technologies for queries on healthcare system stress across the U.S. Our specific aim is to configure open-source tools for complete and rapid data collection and reporting of system stress.

Methods
REDCap is a secure, web based data collection application used to build and manage online surveys, which can be sent to participants by email, and data collection instruments, which require a REDCap user account\textsuperscript{1}. Piwik is an open-source web analytics platform that provides weekly reports on REDCap usage. We defined specific requirements for our project and how we intended to use the open-source software solutions. These requirements were evaluated against the existing feature in each open-source tool and we reached out to user-communities to better understand existing solutions and configurations, as needed. In the instances that the open-source tools could not achieve our specific requirements we leveraged open APIs to configure our own solution. The infrastructure is developed to meet requirements for rapid content configuration and deployment to a large database of participants within 24 hours. A tool that meets these requirements could be used for non-emergent events as well.

Results
The query on healthcare system stress was designed as a REDCap survey and 354 participants’ names, emails, unique identifiers, and responses were managed via the Participant List REDCap feature. By inputting a unique identifier we were able to pre-populate survey responses using information participants provided previously, track participants’ responses, and send reminder emails (see Figure 1). We sent 289 automated reminder emails to participants who had yet to respond and observed an increased rate of responses per hour: 4 responses per hour compared to 1 response per hour prior to reminders being sent. We identified the requirement to represent geographical distribution of participants visually and developed a geographical visualization tool integrated with REDCap. The visualization tool uses real-time data and overlays responses with population density maps and creates the 3 color-coded maps necessary to represent participation. The analytics platform was used to track website visits (n = 178) and length of each visit (average 498 seconds). We additionally collected data on participant actions performed, participant city, device type, and Internet browser.

Conclusion
We successfully fulfilled all of our project requirements except for an ideal solution to pre-populate fields using demographic participant list data. Our REDCap survey configuration is limited to surveys and requires a consistent participant identifier across all projects on a server. More configurations are necessary to find an efficient way of pre-populating additional field types, not limited to text boxes, at the project level instead of at the server level. In comparison we were able to easily find open source tools that we could configure for our usage analytic needs.

Acknowledgements
Funding by the Office of the Assistant Secretary for Preparedness and Response (ASPR)

References
Improving Pain Assessment in Medical Intensive Care Unit Through Natural Language Processing

Douglas Redd, PhD¹,²,⁵, Cynthia Brandt, MD, MPH,³,⁴, Kathleen Akgün, MD, MS³,⁴, Jinqiu Kuang, MS¹,⁵, Qing Zeng-Treitler, PhD¹,²
¹VA Salt Lake City Health Care System; ²Medical Informatics Center, George Washington University; ³Department of Medicine, Section of General Internal Medicine, VA Connecticut Healthcare System; ⁴Department of Medicine, Section of General Internal Medicine, Yale University School of Medicine, New Haven, Connecticut; ⁵Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Introduction
Medical intensive care unit (MICU) admissions are increasing for older adults, especially towards the end of life (EOL). Among MICU patients with a high risk of death, pain is reported by up to 80% of patients and is often moderate to severe in intensity. Inadequate symptom management is consistently identified as an area for improvement amongst MICU survivors as well as family members.

Sensitive, validated tools for measuring pain are readily available for use in the MICU. These tools can be used to enter pain assessments into the electronic health record (EHR) as structured or unstructured data. Inpatient pain assessments are routinely entered as unstructured data. Development of future interventions to improve the quality of pain assessment and management depends on developing an accurate approach to extracting pain assessments from unstructured MICU data. We are developing a pain score extractor for this purpose.

Background
As is common practice, pain assessment in our collection of MICU data is measured on a scale of 0 to 10. In this collection, however, an extra code of 99 is included to indicate that a score could not be obtained (e.g. when a patient is unable to respond). A commonly used method for information extraction of this type of data is the use of regular expressions. Creation and maintenance of collections of regular expressions can be cumbersome and brittle, and requires specialized knowledge of the regular expression language. For this reason, we employ the Regular Expression Discovery for Extraction (REDEX) algorithm, a supervised learning algorithm for automated discovery of regular expressions. This approach has been used successfully in other studies for extracting temporal expressions, body weight, and METs values.

Methods
Documents containing the keywords “pain” and “score” were retrieved using the Voogo application within the VA Informatics and Computing Infrastructure (VINCI) system. These documents were split into snippets (the matched keyword and the surrounding context), resulting in 3862 snippets. These snippets were human annotated to indicate pain scores. A REDEx model was trained using these snippets, with precision .90, recall 0.94, specificity 0.90, f-measure 0.92, and accuracy 0.92. This model was then applied to a random sample of 1000 MICU unstructured documents from an approximation of the VA Musculoskeletal Diagnosis (MSD) cohort. Pain scores were extracted from these documents using the REDEx model. Pain scores were also queried from structured data for the same patients and the same date as the unstructured documents for comparison.

Results & Discussion
815 pain scores were extracted from the unstructured data, while 4063 pain scores were available from structured data for the same MICU patients at the same time. This represents a 20% increase in the number of pain score observations over structured data alone. This demonstrates that the range and representation of pain scores can be substantially increased by including those extracted from unstructured notes in the development of future interventions.
Event Coreference in Support of Temporal Reasoning in Mental Health Notes
Ruth M. Reeves, PhD1,2; Marcus Verhagen, PhD3; Cynthia A. Brandt, MD, MPH4,5; Wendy W. Chapman, PhD6,7; Michael E. Matheny, MD, MS, MPH1,2; Steven H. Brown, MD, MS1,2; Brian Marx, PhD3,8; Theodore Speroff, PhD1,2

1VA Tennessee Valley Healthcare System, Nashville, TN; 2Vanderbilt University, Nashville, TN; 3Brandeis University, Waltham, MA; 4West Haven VAMC, West Haven, CT; 5Yale University School of Medicine, New Haven, CT; 6Salt Lake City VAMC, Salt Lake City, UT; 7University of Utah School of Medicine, Salt Lake City, UT; 8VA National Center for PTSD, Boston, MA; 9Boston University School of Medicine, Boston, MA

Introduction: As part of the development of an NLP engine to assist with clinical reasoning among mental health documents of patients in care for post-traumatic stress disorder (PTSD), we tested the hypothesis that conceptual relatedness between events mentioned in narrative documents can be used to recognize mentions referring to the same event. For example, the mention “sudden anxiety” occurring in one document of a patient may be a reference to the event described as “panic attack” in another document. The circumstances under which such pairs do and do not have the same referent are examined in this work. Establishing co-reference between event mentions within a single document, or across documents of a given patient is a central method within the temporal reasoning NLP pipeline, MedTARSQI developed for this project. Merging event information between co-referring event mentions provides opportunity to assign more precise temporal information to all instances of the same event, and to other event members of temporal relations (e.g., BEFORE E2) within an event sequence. With each iteration of event-merging between co-referring event mentions, and consequent temporal relation assignment along each event sequence, a greater number of events can be anchored by temporal values. This in turn results in a decrease in interval lengths that constrain the temporal interpretation of events within a sequence that are not anchored by temporal values. That is, we have a more precise temporal interpretation of Event, in the relation [E1 BEFORE E2] if we know that Event1 occurred at a given time-date. This submission lays the groundwork for identifying co-referring events as candidates for merging temporal information of distinct event mentions.

Methods: From a cohort of over 3,000 veterans receiving care for PTSD, we selected a random set of 58 patients having at least 3 mental health narrative notes. A corpus of 368 mental healthcare notes from this selection was used to conduct an event co-reference detection test. To detect events in the corpus, we ran YTEX, a searchable database tool underwritten by the NLP engine cTAKES. We filtered the terms detected by YTEX by matching to spans of text that were annotated for PTSD symptoms and treatments. The database functionality of YTEX was used to group event mentions within each patient dataset by the concept code within the UMLS that each mapped to. Within each patient dataset, terms recognized by YTEX which mapped to the same concept code, or a daughter concept code were placed in a “synonym set” as coreference candidates. These terms were then assigned an event profile feature set composed of an event classification, an association with a temporal expression, and an event domain. Event pairs within each synonym set were determined to be co-referent if and only if their event profiles contained the same values for each feature; otherwise, the pair was determined non-referent. We compared the co-referent/ non-referent assignment of each pair in the synonym sets for 12 patients in the study sample against human review of the sets.

Results: YTEX identified a total of 6,820 token strings, mapping to 154,017 possible concepts from the mental healthcare note corpus (368 notes). Forty-eight patient datasets had terms forming synonym sets, with a total of 5,736 concepts determined to be events. Pairs of events from 12 patient datasets were reviewed as referent or not co-referent and compared against the assignment of synonym-set-with-constraints, with a rate of 74.2% agreement, where agreement is a binary measure for each event pair presented: co-referent or not co-referent.

Discussion: In conducting the human review of candidate event pairs, we spotted pairs that were wrongly determined to be co-referring by the features in the event profile. One error type stemmed from the failure to distinguish hypothetical from experienced events; another resulted from events with distinct participants, as occurring with trauma exposure caused by witnessing as opposed to experiencing. These error types could be corrected with a finer-grained event classification system.

Limitations: The cohort was selected by numbers of notes, rather than distributing across a demographically representative group. This study does not account for the proportion of event pairs that were not recognized as possible coreferences (False Negative pairs) where one member of the pair was not identified as an event.

Conclusion: Using event recognition and conceptual relatedness among event terms can reliably determine coreference relations with a small set of features to constrain the assignment. This aids in achieving event sequentialization among narrative notes.

References
1. Reeves, RM; Ong, FR; Matheny, ME; Denny, JC; Aronsky, D; Gobbel, GT; Montella, D; Speroff, T; Brown, SH; Detecting Temporal Expressions in Medical Narratives; IJMI, Feb 2013, Pp 118-127

This research has been reviewed and approved by the Tennessee Valley Healthcare System IRB at the Nashville VA facility. The material is based upon work supported by the Veterans Health Administration, Office of Health Services Research and Development and with resources and the use of facilities at the VA Tennessee Valley Healthcare System. Full source: VA HSRD IIR 12-362.
**Pediatric Inpatient Burden and Comorbidities of Bullous Pemphigoid in the United States**

Ziyou Ren, BS¹,², Jonathan I. Silverberg, MD, PhD, MPH¹

¹ Department of Dermatology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611; ² The Graduate School, Northwestern University, Chicago, IL 60611.

**Introduction**

Bullous pemphigoid (BP) is an autoimmune blistering disorder associated with substantial morbidity. The incidence and prevalence of BP have been reported to be between 0.2 and 3 per 100,000 person years.¹ Previous studies showed that BP was more common in older age groups and associated with female sex.²,³ Analyses based on National Inpatient Data showed that patients with a primary (mean ± std error: 75.9±0.3 yr) diagnosis of BP were significantly older than those without a diagnosis of BP (57.0±0.1 yr); 87.7% of with a primary diagnosis of BP were age 60 years or older. Therefore, few large-scale epidemiological studies have examined risk factors for BP for the pediatric patients. We hypothesized that pediatric patients with BP would have higher rates of comorbid neurologic disorders, as well as autoimmune, infectious and cardiovascular disorders than adults patients. We further hypothesized that the combined disability of BP and associated comorbidities would contribute towards higher rates of hospitalization in pediatric patients with BP compared to those without BP.

**Method**

The 2002 – 2012 Nationwide Inpatient Sample (NIS) provided by the Healthcare Cost and Utilization Project (HCUP) from the Agency for Healthcare Research and Quality (AHRQ) was analyzed. The databases were searched for a primary and/or secondary diagnosis of BP using ICD-9-CM codes and were categorized by age. Univariate and multivariate linear regression and logistic models were constructed to analyze the risk factors and comorbid disorder association.

**Result**

The total incident rates of BP were 16 for primary and 20 for secondary diagnosis of BP out of 13 million pediatric patients. We found that social conditions as well as hospital conditions were significantly associated with pediatric patients with BP. Pediatric patients in a teaching hospital would have a (odds ratios [95% confident interval] 6.2[1.3-30.0]) higher risk of BP than those not. Furthermore, they would have a 0.2[0.07-0.5] lower risk of BP if they came from lower income family than higher income family. For comorbidity profile, pediatric patients with 2-5 and more than 6 chronic diseases experienced 13.5[5.9-31.0] and 40.2 [13.1-123.5] times more hazard than patients with 0-1 chronic diseases. Out of 130 comorbidities tested, 20 showed significant associations with BP in pediatric patients. Specifically, pediatric patients would be exposed to a higher risk with hypertension 63.4[14.6-275.4], deficiency anemias 17.1[7.2-40.5] and bacterial infection 14.5[6.1-34.5].

**Conclusion**

There appeared to be economic and healthcare disparities with respect to BP, such that pediatric patients from rich family appeared to have higher odds of hospitalization of BP in teaching hospitals. Furthermore, pediatric patients with BP experienced much higher risks of autoimmune, infectious and cardiovascular comorbidities than both pediatric patients without BP and adult patients with BP.

**References**


Development of Mobile Application System to Detect Pathological Smartphone Use and Smartphone Addiction

Mi Jung Rho¹,², In Hye Yook¹, Heon Baek¹, Mun Joo Choi¹, Seo Joon Lee³, Dai Jin Kim⁴,⁵, In Young Choi¹,²*

Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea¹
Catholic Institute for Healthcare Management and Graduate School of Healthcare Management and Policy, The Catholic University of Korea, Seoul, Republic of Korea²
BK21PLUS, Department of Health Policy & Management, Graduate School, Korea University³
Addiction Research Institute, Department of Psychiatry, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea⁴
Department of Psychiatry, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea⁵

*Corresponding author: iychoi@catholic.ac.kr

Backgrounds and aims: The number of Smartphone users is increasing in worldwide and 40 million people (78.6%) have Smartphone in South Korea. The explosive usage of Smartphone leads negative results such as Smartphone addiction. This study aims to explore the feasibility of screening and diagnosis of pathological Smartphone use and addiction using mobile technology.

Methods: We developed smart healthcare platform for Smartphone addiction patients (MindsCare system). The MindsCare system was developed to gain Smartphone use pattern to prevent excessive use. It was developed using open source software and calculated time of usage and data for power consumption including Central Processing Unit (CPU), Wireless fidelity (Wi-Fi)/Cellular data, and Liquid Crystal Display (LCD). We collected data from 24 people who agreed to participate in the experiments. They were allocated to the addiction group and control group by using Korean Smartphone addiction proneness scale for adults. We collected time spent for smart phone applications for four weeks and analyzed the Smartphone use pattern by application categories.

Results: Average Smartphone usage time is 3.5 hours for addiction group and 1.3 hours for control group. The mostly used application category for addiction group and control group is Social Network Service (SNS), entertainment, game and web browser. The daily average time for SNS is different as 33 minutes for addiction group and 19.6 minutes for control group. Entertainment is 13.5 minutes for addiction group and 9.7 minutes for control group.

Conclusions: The mobile application system is useful for screening and diagnosis of pathological Smartphone use and addiction. Further development and research will be conducted to improve the effectiveness of the mobile application system.
PubMed ‘Early Alerts’: Towards Better Precision of Literature Searching for Pharmacovigilance Information Based on an Assessment of Relevance Feedback

Anna Ripple, MLS¹; Alfred Sorbello, DO, MPH²; Shahrukh Haider, DDS, MPH²; Olivier Bodenreider MD, PhD¹

¹Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA; ²US Food and Drug Administration, Center for Drug Evaluation and Research, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

**Background:** The PubMed ‘Early Alerts’ provide FDA regulatory reviewers with weekly topical searches of the most recently submitted citations to PubMed/MEDLINE to support prospective detection of emerging adverse drug events for specific drugs. We seek to increase the precision of electronic searching based on an assessment of relevance feedback for a subset of retrieved citations for the antidiabetic medications.

**Methods:** Using a search query optimized for recall and focusing on the titles, abstracts, and keywords, four regulatory evaluators assessed 30 citations each for relevance to drug safety and efficacy drawn from a random sample of 120 of the most recently deposited reports (i.e., not yet indexed). Candidate precision terms were identified by significant word frequency analysis (excluding stopwords) using Word Counter Tool, a free online word counter. We performed a differential frequency analysis of text word occurrences for the relevant compared to non-relevant citations.

**Results:** Based on reviewers’ feedback, half of the reports in the 120 citation random sample were assessed as relevant to safety, efficacy, or both safety and efficacy, whereas the remainder were considered non-relevant for both safety and efficacy. The candidate precision words ‘efficacy’, ‘safety’, ‘adverse’, and ‘risk’ were identified by expert opinion from the titles or abstracts of the relevant citations among frequently occurring words (see Table 1 below). The text words ‘efficacy’, ‘safety’, and ‘risk’ had no title occurrences in the non-relevant citations. However, ‘risk’ and ‘safety’ are not discriminant, because they also occurred in the abstracts of non-relevant citations, leaving ‘adverse’ as the only candidate precision word derived from the abstract field. Analysis of the keywords field did not yield any candidate precision text words.

<table>
<thead>
<tr>
<th>Relevance Assessment</th>
<th>Citation Count, n/N (%)</th>
<th>Text Word Field and Candidate Precision Words identified from Citations in each subset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Relevant citations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-relevant citations</td>
</tr>
<tr>
<td>Safety only</td>
<td>38/120 (32%)</td>
<td>Title: efficacy; safety; risk Abstract: safety; adverse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract: risk</td>
</tr>
<tr>
<td>Efficacy only</td>
<td>49/120 (41%)</td>
<td>Title: efficacy; safety Abstract: safety</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract: risk</td>
</tr>
<tr>
<td>Safety and efficacy</td>
<td>27/120 (23%)</td>
<td>Title: efficacy; safety Abstract: safety; adverse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Abstract: safety</td>
</tr>
</tbody>
</table>

**Conclusions:** We identified several candidate text words to potentially increase the precision of our literature search strategy for pharmacovigilance information, but further validation is required. Small citation sample size and restriction to only the title, abstract, and keywords fields may have contributed to the limited capacity to identify such candidate precision words. Future research is planned involving a larger group of citations, a broader set of drugs and the application of more sophisticated analytical techniques such as term frequency-inverse document frequency to detect text words that may enhance the precision of the PubMed ‘Early Alerts’ as a tool to complement other ongoing pharmacovigilance activities at FDA.

**Acknowledgements:** Funding support received from the FDA/CDER/OTS and the Intramural Research Program, NIH, National Library of Medicine. **Disclaimer:** The views expressed are those of the authors and do not necessarily represent the views of the US FDA, the NIH, or the US Government.
Use of a market segmentation method to validate user types of SHUTi, an Internet-delivered intervention for insomnia.

Authors:, Ritterband, L, PhD¹, Sturz, V, MS², Chow. P., PhD³, Lord, H., MS¹, Thorndike, F., PhD¹, Cohn, W, PhD².

¹Dept. of Psychiatry and Neurobehavioral Sciences, Center for Behavioral Health & Technology
²Dept. of Public Health Sciences
University of Virginia, Charlottesville, Virginia

Background: Given increased barriers to accessing health care, Internet-delivered interventions have increasingly become an option in the delivery of health services, particularly in behavioral health. Hundreds of these web-based treatment programs exist in various stages of development, evaluation, commercialization, or implementation. These interventions are typically defined as self-guided, tailored, interactive, and engaging interventions modeled after effective face-to-face treatments but provided via the Internet. Unfortunately, most commercially available programs have little scientific basis or empirical validity, and even fewer have been evaluated to understand who might be best served by this type of approach.

Methodology: This is the first study to use the Tailored Educational Approaches for Consumer Health (TEACH) model¹ to segment users of an Internet intervention for insomnia (Sleep Healthy Using The Internet or SHUTi)² based on users’ personal characteristics and preferences for health information. Participants were administered the TEACH measure at baseline as part of a large RCT testing the efficacy of the insomnia intervention in comparison to online patient education in a population seeking help online for chronic insomnia. The TEACH measure was given to determine each participant’s strongest affiliation with one of the eight previously validated segments

Evaluation Results: SHUTi participants (n=151) were primarily affiliated with two segments, Segment 4: “Unhealthy, Ask-the-doctor” (72.2%) and Segment 1: “Health Conscious, Browsers” (20.5%). Very few participants were categorized as affiliated with the other segments (7.3%). Segment 4 is characterized as needing information that is credible, available via a healthcare provider and requiring less reading. Segment 1 is characterized as needing information delivered by the Internet provided at a higher reading level and comprehensively. Both of these segments are reasonably expected to seek out and engage in health interventions, including online interventions. As reported previously, participants achieved substantial clinical health benefit³ suggesting that Segments 1 and 4 can benefit from an Internet-delivered intervention.

Conclusions: This data confirms that there are certain types of people who seek out and engage in self-help digital health interventions more often than other user types. In conjunction with previously presented outcome data³, this data highlights that these types of users can perform well in an online health intervention, but future research needs to evaluate both how to better recruit individuals who fall in other segments and are thus under-represented as well as how to adapt interventions to meet a diverse set of users. Finally, new research must test whether individuals in other segments, once recruited, can achieve the same health benefits as their peers in Segments 1 and 4, or whether only certain segments do well with digital health interventions.

References:
Improving Inclusion of Comparison Studies in Radiology Interpretations

Catherine C. Roberts, M.D.¹, Shirley A. Loprino, R.T.(R)¹, Elaine M. Comstock, R.T.(R)(CT)¹
¹Mayo Clinic, Scottsdale, AZ

Introduction
A comprehensive satisfaction survey of physicians that refer patients for imaging at our institution revealed an opportunity for improvement regarding the use of prior comparison studies in the interpretation of new studies. This feedback about the suboptimal use of comparison studies was received for all radiology subspecialties in our practice (ex. neuroradiology, chest radiology). Comparing with prior imaging is a very useful adjunct in assessing the chronicity of imaging findings and thus limiting the provided differential diagnosis. This has a direct impact on patient care since broad differential diagnoses, as well as indeterminate findings, typically warrant additional imaging or clinical investigation, some of which may be avoided. Failing to compare radiological studies to prior studies can cause unneeded procedures and additional imaging for patients, exposes the institution to legal risk, and can delay appropriate patient care.

Methods
The target was to have > 80% of radiology interpretations at our institution either compare to a prior study when available or document that there were no prior studies for comparison. This was an arbitrary goal that left opportunity for physician discretion to not compare if they did not feel it would add value. The radiology imaging electronic medical record was searched using the Softek Illuminate InSite program to establish the rate of documentation of comparison studies in all types of radiology reports. All radiology reports in the designated time period were searched for containing any form of the word compare (comparison, comparisons, compare, compared, compares), prior(s), change or since. Reports using the word “since” were manually reviewed to confirm that the word was used in reference to comparing with a prior exam.

Results
At baseline, 63% (13,421 out of 21,216) of radiology reports generated in a single month included some reference to a prior study. The denominator was the number of reports containing the trigger words. There were no false positives. Numerous factors contributing to the gap in quality of care were identified and addressed. Four Plan-Do-Study-Act (PDSA) quality improvement cycles were undertaken that implemented a number of changes, which included alterations in workflow, automated image retrieval, and generation of voice recognition software templates. Of note, the first intervention of simply informing the different radiology divisions of their performance was entirely unsuccessful in producing change. After the fourth PDSA cycle, compliance increased to 84.7% (14,638/17,273).

Since searching for, retrieving, and reviewing prior studies takes time to complete, our countermeasure was that our report turnaround time (time from images available to report finalized) would not increase by more than 10%. Having radiology reports rapidly available to referring physician is a high satisfier for the referring physicians and is, of course, very important for timely patient care. We found no significant change (p-value = 0.71) between the average report turnaround times at baseline (129 minutes) and at the completion of the project (115 minutes), indicating no adverse effect of the project interventions to increase the use of comparison radiographic studies in interpretations.

Conclusion
Our radiologists have a tremendous amount of pride in their work. It can be difficult to ask people to change, since this is often taken as a criticism regarding their prior work. It is of utmost importance to approach changes such as these with a receptive customer-centered attitude. A combination of behavioral and technologic changes in workflow led to a successful project to improve the quality of our patient care.
Developing Standards for Usability in Electronic Health Records: Challenges and Opportunities

Mitra A. Rocca, Dipl. Inform. Med1, Rebecca Meehan, Ph.D.2, John Ritter, MSc.3, Gary Dickinson4, Kandace Kelly5, Lauralee Barrett6, Ethan Budreau7, Constance Johnson8, Ph.D., FAAN

1Food and Drug Administration, Silver Spring, MD; 2Kent State University, Kent, OH; 3Health Level Seven, Ann Harbor, Michigan; 4CentriHealth, Los Angeles, CA; 5 6 7EPIC, Verona, WI; 8Duke University, Durham, NC

Abstract

The Health Level Seven (HL7) Electronic Health Record System (EHR-S) usability project is an ambitious collaborative effort that addresses the importance of usability in EHR-S and the challenges of developing standards. The goal of this project is to translate existing, well established usability guidelines and health information management principles into functional criteria for EHR-S.

Introduction

Views of what usability is, and how to improve EHR usability vary across stakeholder groups. Usability is defined as how useful, usable, and satisfying a system is for the intended users to accomplish goals in the work domain by performing certain sequences of tasks1. There have been many advances made to improve EHR usability, yet, they are often implemented in silos or shared in a fragmented way, preventing feedback from different stakeholders. This poster will present the collaborative effort of the HL7 EHR-S Usability work group to highlight the importance of EHR usability standards, the process of developing standards to improve EHR usability and the approach for implementation. The HL7 EHR-S Usability workgroup focuses on translating existing, well established usability guidelines and health information management principles into functional criteria complete and develop two functional profiles: 1. User Centered Design Functional Profile 2. Systems Usability Design.

Methods

EHR usability is one of the most important issues influencing safety outcomes of patients in today’s healthcare environment. In order to achieve the project goals, the work group:

1) Conducted an environmental scan by reviewing the literature and government agencies guidelines (e.g. NISTR 78042, ONC SHARP-C3, HIMSS EHR Usability Taskforce4).
2) Using the NISTIR 7804 report as a basic guide, the workgroup compared the heuristics already developed to determine any missing heuristics within the NISTIR 7804.
3) Developed the EHR usability heuristics and framework that can result in an EHR-S Usability functional profile complete with functions, conformance criteria and examples for the HL7 balloting process.

Results

The work group members addressed the comments resulted from the HL7 ballot process and plan to ballot the revised EHR usability functional profile as an informative ballot in the HL7 September 2016 ballot cycle.

Conclusion

Usability standards need to be a critical aspect of EHR user interface design to ensure that healthcare providers can use these systems in an effective, efficient, safe and satisfying manner.

Future Plans

The HL7 EHR-S Usability work group aims at providing the developed standard for accrediting EHR systems.

References

Are Boston Healthcare Center Websites Linguistically Accessible?

Jorge A. Rodriguez, MD, Sanja Percac-Lima, MD, PhD
Massachusetts General Hospital, 55 Fruit St., Boston, MA 02114

Introduction
The Internet has become a significant source of health information, with nearly 87% of Americans having access to the Internet1. The availability of linguistically appropriate content is critical to providing patients with access to health information. While accessible health content is important for English-speaking patients, it becomes paramount for Limited English Proficient (LEP) patients. Much like the health disparities among racial, educational, and economical lines, language disparities have been associated with negative health outcomes2. The LEP population is growing, with 20% of the US population speaking a language other than English at home and 8.6% defined as being LEP. In Boston, 12% of the population is LEP3. Furthermore, homepages are the online face of healthcare centers and access points for all patients4. Thus, with this context in mind, we sought to evaluate the language accessibility of healthcare center websites in Boston.

Methods
We reviewed the homepages of Boston hospitals and community healthcare centers to identify if there was an option to translate the homepage, and if so, the type of translation method provided and the number of available languages. We categorized the translations as automated (i.e. Google Translate), single translated homepage, and entire separate translated website. After we categorized each health center, we mapped their translation capacity to the percent LEP population in their neighborhood.

Results
We reviewed 45 homepages, 20 of which were hospital websites and 25 were community health center websites. Only 44% of homepages had a translation option available. When we looked at the type of translation methods, we found that 60% of the homepages had automated translation services, primarily Google translate, 15% had a single translated homepage in specific languages, and only the Dana-Farber had a separate translated website in Spanish. The number of available languages ranged from 1 to 83. Among the 10 neighborhoods with >33% LEP population, only 50% of the healthcare centers had a translation option available.

Conclusion
Hospital websites serve as access points for the communities they serve. However, despite the linguistic diversity of Boston neighborhoods, many of the hospitals and community centers do not provide translated versions of their websites. With the shift of health information to the Internet, the inability to access healthcare center websites serves to further extend health disparities that already exist for LEP patients. We plan on extending this type of review to the state of MA and other major cities in the US.

References:
Towards automatic discovery of Genes related to Human Placenta

Laritza M. Rodriguez, MD, PhD, Stephanie M. Morrison, MPH, Kathleen Greenberg, PhD, Dina Demner Fushman, MD, PhD

Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction

The human placenta is the most important organ in human pregnancy. Its functionality influences maternal and fetal health and development and impacts future health for both mother and baby1. With the advancement of genetic and molecular sciences, it has become possible to study placental biology during pregnancy.

Discovery of gene pathways and biochemical mechanisms that help explain disease causality can be facilitated by automated extraction of relevant information from the biomedical literature2. One essential step in understanding complex pathways has been the compilation of disease-specific gene candidates extracted from the literature. Data analysis then makes it possible to create candidate gene assays and translate the results to the wet bench for biochemical research.

Our goal is to extract genes, gene pathways, biomarkers, and related events from the human placenta literature to create a specialized human placenta gene repository, and to identify pathways that can uncover target genes and gene therapies for pregnancy-related diseases. Here we present the first phase of the study: extraction of gene mentions from text.

Methods

The document collection was retrieved from PubMed® using search filters and the following search terms: placenta, gene, biomarker, polymorphism, enzyme, preeclampsia, hypertension, diabetes, growth restriction. The tailored search returned 428 papers. To identify specific gene mentions from titles and abstracts we used the NLM MetaMap 20143 restricting processing to semantic type Gene. We then manually classified the extracted gene names into: 1) specific gene mentions (e.g. CDKN1A), 2) general genetic terms (e.g. Growth Factor Gene), and 3) errors (e.g. PREECLAMPSIA, SUSCEPTIBILITY TO).

Results

In the 428 retrieved abstracts, MetaMap identified 413 with Gene mentions. Overall, MetaMap identified 826 distinct gene names in titles and abstracts in the collection. Of these, we classified 753 as specific gene mentions, 71 as general genetic terms, and 2 as errors. Examples of the most frequent mentions by type are: Specific Gene: NLRP5 gene (178), NCR3 wt allele (83), CD8A wt allele (83), CD69 wt allele (71), PGF gene (26), LEP gene (18); General genetic terms: Genes (246), Alleles (51), Genome (37), Human gene (34); Errors: Genes, vif (50), PREECLAMPSIA, SUSCEPTIBILITY TO (2).

Conclusions

Our preliminary analysis shows that the placenta literature contains enough specific gene mentions to warrant further text mining on the genes of interest to identify pathways, biomarkers, and relationships between placenta gene expression and maternal and/or fetal diseases, ultimately identifying predictors of diseases that may clinically manifest only later in life.

References


Family History by Other Names: A Preliminary Comparison of Structured and Free-Text Sources in the Electronic Health Record

Paul T. Rosenau, MD, MS¹, Diantha B. Howard MA, MS¹, Genevieve B. Melton, MD, PhD², Elizabeth S. Chen, PhD³
¹University of Vermont College of Medicine, Burlington, VT; ²University of Minnesota, Minneapolis, MN; ³Brown University, Providence, RI

Introduction

Family history (FH) data documented in electronic health record (EHR) systems have the potential to inform studies of heritable determinants of health and disease. A unique aspect of pediatric documentation is its common reference to birth history, a potentially rich source of maternal FH data¹. We sought to characterize structured and free-text sources of FH data in the enterprise EHR used at the University of Vermont Children’s Hospital (Epic Systems Corporation, Verona, WI) with respect to variability and redundancy of FH documentation.

Methods

We compared FH data from two sources: 1) the FH section of the History ‘Activity’ (“FH Activity”), a set of updatable, structured fields accessible from a variety of clinical contexts, and 2) free-text History and Physical (H&P) notes that document inpatient admissions. Records of patients born January-March 2010 who were admitted to the hospital at least three times before their fifth birthday were considered eligible. FH data were manually abstracted from all H&P notes signed by an attending physician within 48 hours of admission into pairs of discrete observations (e.g., presence of a disease) about family member relations. Observation-relation pairs, or “statements”, were compared to the FH Activity data with respect to novelty; concept granularity; and frequency of relations and clinical topics referenced by the statements. The relative novelty of statements were scored in terms of their granularity (general vs. specific), scope (inclusion or absence of associated attributes), and semantic vs. syntactic equivalence. Patient-specific statements were compared across encounters and against FH Activity data available at the time of the patient’s last hospital encounter. Statements were scored as partially-novel if they differed from all other patient-specific statements and FH Activity data by at least one dimension and this difference produced a net increase in granularity or scope specification.

Results

Fourteen patients met inclusion criteria, contributing twelve sets of FH Activity data and 68 notes from 53 hospital admissions. FH Activities and H&P notes contained 50 FH entries and 459 statements, respectively. All notes contained statements except for four of the five surgical H&P notes. Frequency of statements varied by hospital unit. Admissions to the neonatal intensive care unit (NICU) generated by far the highest number of statements per patient admission (13.1±7.3.), a result of note duplication. The proportion of partial novelty ranged from 84% in the newborn nursery (NN), to 19% in the pediatric intensive care unit (PICU). The general topics covered by FH statements also varied by unit. For example, common PICU subjects were Cardiovascular (59.3%), Gastrointestinal (14.8%), Statements of Uncertainty (14.8%), and Infectious Disease (7.4%). In contrast, the top NN statements pertained to Infectious Disease (43.6%), Mental Health (14.1%), Hematological (11.5%), and Endocrine (9.0%). Female relations were noted to be much more frequent than males. For newborn admissions the female to male relations ratio was 72:1; for non-newborn admissions, this ratio was 1.4:1.

Discussion

In this small dataset, H&P notes contained unique FH data the frequency, novelty and subject of which varied by hospital unit. Birth history data found in birth admission notes were a particularly rich source of maternal FH. This study represents an initial step toward characterizing a pediatric FH dataset from a commercial EHR in an academic health center using a methods that may be generalizable to other EHR systems and institutions.

Acknowledgment: This work was supported in part by National Library of Medicine grant R01LM011364.

References

1. Abhyankar S, Demner-Fushman D. A simple method to extract key maternal data from neonatal clinical notes. AMIA Annu Symp Proc. 2013 Nov 16;2013:2-9
Patient Record Linkage Between Two Academic Health Centers

Mindy K. Ross, MD\(^1\), Douglas S. Bell, MD\(^1\)

\(^1\)University of California Los Angeles, Los Angeles, Ca USA

Introduction

Clinical research requires complete data on patient treatment and outcomes. However, many patients receive their care from multiple provider organizations and fragmentation of care can adversely impact research on this data. The Los Angeles Data Resource (LADR) is a growing regional collaborative of six health systems in Los Angeles, California that was created to facilitate multi-center clinical and comparative effectiveness research. In order to accurately identify patient cohorts for these applications, it is important to connect disjointed patient records between these healthcare systems.

Eventually, we aim to match patients across institutions utilizing private record linkage, which obviates the need to release patient-identifying information. An example of this is the application developed at Northwestern University, Distributed Common Identity for the Integration of Regional Health Data (DCIFIRHD)(1) using a “one-way hash” algorithm. The accuracy of private record linkage algorithms across systems can be measured by comparison with manually reviewed charts serving as the “gold standard.” However, manual review is resource demanding, which limits the population size for validation studies. The Epic Care Everywhere module imports records on a patient from other health care organizations if a matching patient can be identified at the other institution. We seek to assess the potential to use patient linkages created in Care Everywhere between the UCLA and Cedars-Sinai Health Systems as a gold standard for evaluating alternative private record linkage algorithms.

Methods

We conducted a preliminary exploration of the patient matching data available from the Epic Care Everywhere implementations at UCLA and Cedars-Sinai Health System. At Cedars-Sinai, Care Everywhere is configured to automatically query nearby institutions for patient matches at the time of patient appointments; at UCLA, Care Everywhere queries need to be explicitly triggered.

Results

From November 2014 until September 2015, there were 274,723 query attempts from Cedars-Sinai to UCLA. During this time period, there were 7,347 queries from UCLA to Cedars-Sinai. Queries from Cedars-Sinai to UCLA reveal 207,630 counts labeled as “unsuccessful” and 4,449 queries that were “unsuccessful” from UCLA to Cedars-Sinai. The reasons listed for “unsuccessful” queries are: no matching patients, multiple high threshold matches, multiple patient identification numbers assigned, multiple low threshold matches found, potential match found but patient has no clinical encounters, potential match found but the demographics match score was too low, patient missing Care Everywhere identification number, patient not participating in Care Everywhere, and query timed out.

Discussion

Patient match results generated by Care Everywhere include multiple indeterminate values. The underlying causes of indeterminate results need to be understood before Care Everywhere matches can serve as a gold standard for evaluating other patient linkage algorithms. The next steps are to determine how many of the queries are unique patient matches, manually review 500 patient queries from each institution, and determine the sensitivity and specificity of the Care Everywhere tool.

Conclusion

We will be able to determine if Epic Systems’s Care Everywhere tool can be used as a surrogate gold standard in place of manual patient query result review.

References

Confidence in Methodologies to Accurately Predict Risk Stratification in Primary Care Practices
Rachel L. Ross, BA, Bhavaya Sachdeva, MPH, Jesse H. Wagner, MA, Lindsey Watson, Jennifer Hall, MPH, David Cameron, Deborah J. Cohen, PhD, David A. Dorr, MD, MS Oregon Health & Science University, Portland, OR, USA

Introduction. Primary care risk stratification is a process where practices apply an algorithm to assign their patients into risk categories based on factors that may contribute to negative health outcomes. Risk stratification (RS) in primary care is crucial to improving outcomes of high-cost, high-needs patients, tailoring interventions for these patients, and controlling healthcare costs. This study will develop a maximum value/minimum complexity approach to RS using a mixed-methods approach. The qualitative arm includes an analysis of online survey responses and follow-up semi-structured interviews. This early analysis of the qualitative work sought to understand the impact of practice characteristics, especially sociotechnical framework elements, on perceptions of RS.

Methods. The survey included questions such as percentage of patients actively risk stratified, changes in RS methodology, type of approach, challenges, and 9 point Likert scale questions to assess success of RS implementation. From a subset of 79 practices that were invited to complete a survey, 65 individual survey responses were received. We assessed these for staff confidence in RS methodologies across practices varying in size, location, and ownership. Data were analyzed by coding as percent positive responses (7-9 on Likert scale) versus other responses, and differences tested by Fisher’s Exact Test. Then, 10 individuals representing 6 practices were interviewed regarding their RS approach. Transcripts were coded using Atlas.ti™ software as relevant themes emerged. Codes relating to the sociotechnical framework were identified, and we analyzed their relationships to other pertinent codes. Correlation coefficients (r) were calculated on how codes co-vary.

Results. As seen in Table 1, our Fisher’s Exact Test revealed that practices’ agreement with the RS scores generated by their algorithm differed significantly by practice ownership, with independent practices more confident, but not by size or location. Independently owned practices are more likely to agree with RS scores and have greater confidence than practices that are part of a system. An analysis of the transcript data demonstrated that quotes coded with “EHR/HIT” co-occurred most frequently with “Risk Stratification Workflow” (c=0.18), “Changing Approach” (c=0.1), and “Computer-based algorithms” (c=0.11). An example of a co-occurrence between “EHR/HIT” and “Risk Stratification Workflow” is the following excerpt: ”I think part of it is the technology. Part of it, on the provider’s side, is encouraging more standardization and approach for a particular group of people...when I came here...a number of these things weren’t fully developed. Though we have been tackling each of those issues since then, you know, human beings are involved so it’s a...[chuckles]...it can be quite a journey, I think, to get there.” Additional quotes, highlighting each co-occurrence, will be provided.

Table 1. Practice Characteristics and Perceptions of Risk Stratification.

<table>
<thead>
<tr>
<th>Ownership</th>
<th>Health System N=36</th>
<th>Independent N=26</th>
<th>Fisher’s exact (p-value)</th>
<th>Chi-Square (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% agree with risk stratification</td>
<td>39%</td>
<td>69%</td>
<td>F=22 (0.0173)</td>
<td>X²= 5.57 (0.0183)</td>
</tr>
<tr>
<td>% confident in risk score accuracy</td>
<td>27%</td>
<td>68%</td>
<td>F=27 (0.0011)</td>
<td>X²=10.75 (0.0010)</td>
</tr>
<tr>
<td>Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large N= 34</td>
<td>Medium N= 23</td>
<td>Small N=8</td>
<td>Chi-Square (p-value)</td>
<td></td>
</tr>
<tr>
<td>% agree with risk stratification</td>
<td>42%</td>
<td>70%</td>
<td>38%</td>
<td>X²= 4.77 (0.0922)</td>
</tr>
<tr>
<td>% confident in risk score accuracy</td>
<td>35%</td>
<td>48%</td>
<td>75%</td>
<td>X²= 4.28 (0.1176)</td>
</tr>
<tr>
<td>Location</td>
<td>Rural N= 18</td>
<td>Suburban N= 18</td>
<td>Urban N= 29</td>
<td>Chi-Square (p-value)</td>
</tr>
<tr>
<td>% agree with risk stratification</td>
<td>50%</td>
<td>31%</td>
<td>64%</td>
<td>X²= 4.48 (0.1067)</td>
</tr>
<tr>
<td>% confident in risk score accuracy</td>
<td>56%</td>
<td>39%</td>
<td>41%</td>
<td>X²= 1.23 (0.5397)</td>
</tr>
</tbody>
</table>

Conclusion. Practices may have more control over their RS process when they are not owned by a system, which could lead to greater agreement with their scoring. Interviewees revealed rich understanding of the complexity of workflow, approach, and technology, consistent with the sociotechnical framework. Further analysis is needed to elucidate the reason for the survey findings and interpret the impact of the co-occurrences found in the qualitative interviews.
Redesigning Chart Rounds in Radiation Oncology

Adam Rule, MS¹, Erin Gillespie, MD², Nadir Weibel, PhD¹, Todd Pawlikci, PhD²

¹Design Lab, UC San Diego, La Jolla, CA; ²School of Medicine, Department of Radiation Medicine and Applied Sciences, UC San Diego, La Jolla, CA

Introduction: Peer review is widely accepted as a useful and necessary form of quality assurance in radiation oncology [1] and many facilities conduct weekly, practice-wide chart rounds to collaboratively review treatment plans for all new patients. However, there is little evidence that current methods of peer review improve quality of care [2]. At a typical chart rounds meeting plans are discussed for an average of just 3 minutes [3], changed only 4-12% of the time [4], and are often reviewed after treatment has begun [5]. In short, chart rounds do not afford rapid and thorough peer review.

Redesigning Peer Review: To redesign peer review in radiation oncology, we assembled a team of radiation oncologists, physicists, and designers at UC San Diego for two half-day workshops. Participants used design thinking [6] to guide the workshops, which encourages thoroughly defining the problem before brainstorming and testing solutions. In the first workshop, participants identified four goals of peer review, mapped how peer review is currently conducted at UC San Diego based on experience and extensive observation, and consolidated observations into “How might we...” statements that captured key areas where peer review could be improved. During the second workshop, participants brainstormed solutions to these design prompts for further prototyping and testing.

Results: Four goals of peer review were identified including:

1. Quality assurance – ensuring patient safety by checking key quality measures of the treatment plan
2. Decision support – helping clinicians make decisions about complex patient cases
3. Education – training residents and senior clinicians on best practices and emerging treatments
4. Team building – fostering a work environment where errors are shared

Participants realized that weekly chart rounds did not provide sufficient time to review quality or appropriateness of all plans and that only certain cases where highly educational. Radiation oncologists at UC San Diego had already limited discussion of routine cases at chart rounds and sought out rapid decision support for difficult cases through mailing lists. Participants recognized these workarounds as an opportunity to provide multi-level peer review and document peer review occurring outside of chart rounds. Participants designed an email review system allowing oncologists to email a treatment plan to a peer for rapid quality checks and decision support. Highly complex or educational cases could then be flagged by the reviewing oncologist for presentation at the next chart rounds. This design seeks to provide peer review for all cases, reduce the time to review, and focus chart rounds meetings on the most profitable cases to discuss.

Acknowledgements: This research was funded by NLM Training Grant #T15LM011271.

References

Leveraging the Electronic Health Record to Investigate Medication Safety Incidents

Alissa L. Russ, PhD\(^1\)\(^-\)\(^3\), Laura G. Militello, MA\(^4\), Peter A. Glassman, MBBS, MSc\(^5\), Karen J. Arthur, PharmD\(^3\), Alan J. Zillich, PharmD\(^1\)\(^,\)\(^2\), Michael Weiner, MD, MPH\(^1\)\(^,\)\(^3\)

\(^1\)Center for Health Information and Communication, Department of Veterans Affairs (VA), Health Services Research and Development Service and Regenstrief Institute, Inc., Indianapolis, IN; \(^2\)College of Pharmacy, Purdue University, West Lafayette, IN; \(^3\)Richard L. Roudebush VA Medical Center, Indianapolis, IN \(^4\)Applied Decision Science, Dayton, OH; \(^5\)VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA.

Introduction
To enhance our understanding of how prescribers and pharmacists recognize and resolve medication safety incidents, we developed a novel approach that integrated cognitive task analysis (CTA) interviews with the electronic health record (EHR). CTA consists of a set of specialized interview and data analysis methods designed to elucidate individuals’ cognitive processes; CTA methods are well-established and help guide informatics designs.\(^1\) This poster describes our methods, the interview guide, and data collection tools.

Methods
We recruited prescribers and pharmacists from a Veterans Affairs Medical Center. Participants reported incidents where they addressed a medication safety problem for any of these categories: 1) adverse drug reaction; 2) drug-drug interaction; and 3) drug-disease interaction involving renal function. Participants submitted incidents via an Incident Card, developed during this study, and incidents within the above categories were eligible for follow-up, CTA interviews. Participants could submit Incident Cards via paper or electronically. The Incident Card captured key data, such as how the participant first identified the problem; resources used to help with decision-making; and actions taken to mitigate safety risks. A physician and pharmacist on the research team reviewed each submission for potential inclusion in subsequent CTA interviews. We integrated CTA with the EHR in a novel manner to expand data capture and used EHR-stimulated recall with participants to aid reconstruction of safety incidents. Interviews were audio recorded for future, qualitative analyses focused on decision-making processes.

Results
We captured 101 incidents; most (79%) were eligible for interviews. After clinical review, we conducted CTA interviews on 60 incidents: 31 submitted by prescribers and 29 by pharmacists. Incidents included 20 adverse drug reactions, 20 drug-drug interactions, and 20 drug-disease cases. Participants consulted the EHR to recount the incident and reconstruct a timeline of events (e.g., days and times when medications were ordered, modified, and discontinued). Participants consulted pharmacy records, progress notes, lab results, allergy documentation, and other EHR data to explain key points that helped them recognize a problem and determine how to respond. In addition, participants accessed records of personal electronic communication that occurred during the course of the incident, to explain statements that influenced their decisions.

Conclusions
To our knowledge, this is the first study to integrate the EHR with in-depth CTA interviews. Incorporating the EHR aided incident selection, enhanced participants’ ability to reconstruct their experience resolving the incident, and expanded data collection. The EHR improved data capture of details that may be difficult for healthcare professionals to recall, such as medication doses and schedules, facilitating data accuracy and completeness. Leveraging the EHR in this manner could enable the use of CTA in a wide variety of health informatics and clinical decision-making studies, and also may help strengthen patient safety root cause analyses.

Reference
The Impact of Nurses’ Graph Literacy and Numeracy on Comprehension of Visualized Feedback Information

David Russell PhD¹, Nicole Onorato, BS¹, Yolanda Barrón-Vayá, MS¹, Jacqueline Merrill, PhD, RN², Dawn W Dowding, PhD, RN¹,²

¹Visiting Nurse Service of New York, NY; ²Columbia University School of Nursing, New York, NY

Introduction

Providing performance feedback to clinicians is an important factor in improving health care quality¹ and is a key component of the IHI Triple Aim initiative². Dashboards, a form of Health Information Technology (HIT) which use visualization techniques to summarize performance against metrics, can be used to provide feedback. The aim of this study is to evaluate the relationship between nurses’ graph literacy and their ability to comprehend feedback information related to the care of heart failure patients, presented in a dashboard format.

Methods

A multi-factorial experimental research design using an online survey. Graph literacy was measured using the graph literacy scale³ which was developed specifically for the health domain and measures graph reading skills and comprehension across different types of graphs. Numeracy was measured using the expanded numeracy scale⁴. 129 nurses at a large non-profit home care agency were randomly allocated to an experimental condition where they received visualized information about specific quality indicators (Table 1). Outcomes include knowledge and understanding of the information presented in the visualized dashboard.

Table 1: Experimental Conditions for the Multifactorial Design

<table>
<thead>
<tr>
<th>Experimental Group</th>
<th>Quality Indicator 1</th>
<th>Quality Indicator 2</th>
<th>Quality Indicator 3</th>
<th>Quality Indicator 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Table</td>
<td>Line Graph</td>
<td>Bar Chart</td>
<td>Spider Graph</td>
</tr>
<tr>
<td>2</td>
<td>Line Graph</td>
<td>Bar Chart</td>
<td>Spider Graph</td>
<td>Table</td>
</tr>
<tr>
<td>3</td>
<td>Bar Chart</td>
<td>Spider Graph</td>
<td>Table</td>
<td>Line Graph</td>
</tr>
<tr>
<td>4</td>
<td>Spider Graph</td>
<td>Table</td>
<td>Line Graph</td>
<td>Bar Chart</td>
</tr>
</tbody>
</table>

Results

Sampled nurses were diverse in terms of their demographic characteristics and years of nursing experience/education. Nurses answered approximately 10 of the 13 graph literacy items and 7 of the 8 numeracy items correctly—slightly higher than average scores for the U.S. population. There were no significant differences across the four experimental groups in demographic characteristics, numeracy scores, or graph literacy abilities. Nurses’ level of graph comprehension varied by the type of display that was presented, adjusting for the type of question. Significantly higher (p<0.05) graph comprehension was observed for nurses who responded to bar graphs compared to spider graphs, and for tables compared to spider graphs.

Conclusion

Results from this study suggest that nurses’ comprehension of visualized feedback information is influenced by the format in which that feedback appears. Efforts are underway to sample additional nurses from another large, regional home care agency. We will examine how graph literacy and numeracy skills are related to graph comprehension. Results will be used to develop flexible dynamic quality dashboards. Future research will evaluate the effectiveness of dashboards in improving care processes and outcomes.

References

MyHealthKeeper: A Pilot study for personal health record based healthcare management

Borim Ryu, MS¹,², Sooyoung Yoo, Ph.D.¹, Eunyoung Heo, MS¹, Sookyoung Ji, MS², Yoojung Kim, MS², Joongsik Lee, Ph.D.², Jeong-Whun Kim, M.D., Ph.D.¹
¹Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, KOREA; ²Seoul National University, Seoul, KOREA

Abstract

Personal health devices (PHDs) are rapidly developing and issuing in medical domain. This personal health record (PHR) based care management systems can be expected to improve patient engagement and data-driven medical diagnosis in clinical setting. In this study, we designed an exploratory user study of collecting data about physical activity and diet using wearable devices in lifestyle related disease patients and clinicians during 8 weeks. The result of our study shows how patients and clinicians accept PHDs. We have successfully demonstrated the user responses and impacts to evaluate the overall usefulness, usability, and satisfaction of PHR-based health management system.

Introduction

Widespread interest in using PHRs has been reported and their adoption has been increasing. Interactive self-management technology with a variety of chronic health conditions is reported to be acceptable and useful. However, few studies have evaluated the effectiveness and value of PHR applications in clinical point of view. In this study, we aimed to demonstrate the clinician evaluations of PHRs using wearable devices in lifestyle related disease patients who has obesity, hypertension, diabetes, hyperlipidemia, or sleep apnea. The ultimate objective of the study is to derive healthcare management service contents and features in a clinical setting.

Methods and Results

We utilized activity tracker (Fitbit, Misfit) to collect individual physical activity data and developed a smartphone applications to record every meals and daily logs. We also monitored data collecting in web administer environment. Patients with lifestyle related disease were participated the study, and asked to use wearable devices for two months. Lifestyle related disease experts were also recruited in the experiment. For patients and clinicians, paper based survey, interview, and workflow observations were processed to monitor and evaluate the study. During the study period, personal health data based clinical diagnosis and recommendations were done in outpatient visit, two times.

The developed diet-log app successfully collected daily meal information, and with the personalized physical activity data, clinicians were able to promote physical activity by the patients. The varied compliance of the participants enabled evaluation of administrative features of the PHR based health management including the generation of participation surveys, and user interviews. We found that data recording compliance and patient expectations to their care providers, as well as clinicians PHR based interventions.

Conclusion

We have developed an innovative testing approach to collect and demonstrate the user interactions of PHRs using wearable devices in lifestyle related disease patients. We have successfully derived the features of PHR service contents and consumer responses, and satisfaction of using wearable devices and PHR-based self-management system.

References

Using patient photos to reduce wrong-patient order entry: Providers’ perspective

Hojjat Salmasian, MD MPH PhD1,2, Jason Adelman, MD MS1,2, Robert A Green, MD MS1,2, David K Vawdrey, PhD1,2
1NewYork-Presbyterian Hospital, New York, NY; 2Columbia University, New York, NY

Introduction

Wrong-patient order entry is an important safety concern that has been estimated to occur in approximately 1/1000 orders placed by prescribers.1,2 Evidence suggests that displaying patients’ photographs in an Electronic Health Record (EHR) system can reduce these errors, and the Office of the National Coordinator for Health IT (ONC) recommends this as a best practice.3,4 However, little is known about the providers’ perspectives of patient photographs being displayed in EHR systems. To assess this question, we conducted a survey of prescribers who were newly introduced to patient photographs displayed in an EHR system to elicit their perceptions of the impact of this process.

Methods

Display of patient photographs was added to an existing patient identification alert for a sample of patients as a pilot project in October 2015. Photographs were taken after obtaining verbal consent, and were displayed when prescribers placed orders for these patients. Prescribers who received identification alerts with photographs were identified using the EHR logs and an anonymous survey was sent to them via email. The survey was delivered using a HIPAA-compliant platform in case any answers accidentally contained patient information, but it was designed in a way that would not ask for any identifiable information about the patients or prescribers. Survey questions focused on prescribers’ demographic information, their experience with the existing patient verification alert, and their perceptions of the new addition of patient photographs displayed within the verification alert.

Results

During the pilot, 57 patients had their photographs taken, and 472 prescribers who cared for these patients were displayed the patient verification alert with photograph. All of these prescribers received the survey, and 103 surveys were completed. Most respondents were resident physicians (43%), nurses (22%) or attending physicians (12%). The majority of respondents reported seeing a patient’s photograph 1 to 5 times (86%). Most (59%) believed the existing patient identification alert without patient photographs was helpful in reducing wrong-patient order entry errors. Additionally, 72% of respondents expressed that adding photographs to this identification alert was at least a moderate improvement, and 25% of prescribers stated this enhanced alert with photographs prevented them from placing orders for the wrong patient. We found no association between the prescriber’s role (nurse, resident, attending) and his or her opinion about the impact of the patient verification alert (Fisher’s p-value = 0.114) or the enhanced alert with patient photographs (p-value = 0.275).

Conclusion

Adding photographs to our existing patient verification alert was perceived positively by residents, nurses and attending physicians, and many prescribers expressed that it prevented them from placing orders for the wrong patient. These findings are consistent with previously published research and a national best practice,3,4 and are informative in the implementation of a hospital-wide process for acquiring and displaying patient photographs. Future research should focus on rigorously quantifying the impact of displaying patient photographs in reducing wrong-patient order entry errors.

References

A Comparative Model of Participatory User Centered Design Methods (pUCD) for Adolescent Health Technologies

Savitha Sangameswaran, MS1, Cynthia M. LeRouge, PhD1, Anne M. Turner, MD, MPH, MLIS1, Torree Malasanos MD2

1University of Washington, Seattle, WA; 2 Kiwee Health, LLC, Gainesville, FL

Introduction
Recent calls from the popular press and academe encourage visionaries, developers, and decision-makers to embrace a human-centered focus by engaging users and stakeholders in the design and selection of health technologies (HIT). Participatory User Centered Design (pUCD) is an approach for gathering information on needs and preferences that can integrate the user’s voice and interaction early and throughout the design process. The UCD domain has evolved into a wide collection of different methods. It is not clear which specific methods under the UCD umbrella are most befitting to health care contexts and with specific groups of targeted users (e.g., adolescents), nor do we have structured guidance on criteria for methodological choice and design in deploying these methods for the development and selection of HIT targeting adolescent users. As digital natives and given the increase in chronic disease in adolescents, understanding which specific pUCD methods are most effective in engaging adolescents is critical for effective design and thereby successful HIT to actualize the promise that HIT may offer this target group. The purpose of this study is to provide structured guidance to researchers and developers on the criteria for methodological choice in deploying pUCD methods for the development and selection of HIT involving adolescent users

Methods
To identify the available methods, we utilized existing toolkits (identified from comprehensive review studies1 and two schools of UCD methodology (Design Thinking and Double Diamond). pUCD Methods were identified and defined with care not to duplicate the same method addressed in the literature under different terminology. The research team then used the Framework for Comparing Assessments of Usability Testing Methods for Children2 (only known UCD framework to target youths) as a lens from which to analyze and develop direction for pUCD methodological choices with adolescents. The model includes the following dimensions: a) Assessment Criteria (e.g., robustness of the method), b) Usability Testing Method Characteristics (e.g., number of participants), and c) Children Characteristics (e.g., trustworthiness of self-report). The purpose of the Framework is not to guide a classification or even a meta-analysis of User Centered Design, but to serve as a way to structure the problem space of selecting appropriate methods. The research team of seasoned pUCD researchers worked collaboratively on assessing analysis and design stage methods (where an understanding of the users’ needs, preferences, and capabilities, as well as creative possibilities, are explored) using the Framework as a guide.

Results
Our resultant Comparative Model of pUCD Methods for Adolescents consists of 22 individual (12 including interviews, user diary, story sharing) and group (10, including focus groups and body mapping) methods. In cases where one activity had distinct activity options (e.g., immersive workshop), we treated each option as a separate pUCD activity. In reviewing these methods, we noted some UCD activities were focused on activities that the design team could do (e.g., brainstorming). In the spirit of pUCD, where possible and practical, we adapted these activities to interactions with potential users. Our results provide structured guidance in the form of a Comparative Model of pUCD Methods for Adolescents

Conclusion
The Framework for Comparing pUCD Methods for Adolescents addresses existing needs by delineating and refining pUCD methodologies in the context of HIT for adolescents to inform and motivate those involved in HIT design, selection and research and move towards a more patient engaged HIT model.

References
Use of a Web-based Survey to Identify Symptom Frequency and Intensity Reporting of Persons Living with HIV with HANA Conditions

Rebecca Schnall, PhD, MPH, RN-BC1, Sabina Hirshfield, PhD2, Karolynn Siegel, PhD3, Lena M. Milian, BSN1, Heidi Castillo, MPhil, BSN, RN1, Hwayoung Cho, MSN, RN1
1Columbia University, School of Nursing, NY, NY; 2Public Health Solutions, NY, NY; 3Columbia University, Mailman School of Public Health, NY, NY

Introduction

HIV has evolved from an acute to a chronic illness largely and as a result, people living with HIV (PLWH) are living longer (1). As PLWH age, they are developing chronic illnesses and co-morbid conditions that are often seen in older HIV negative patients (2). Fifty to sixty percent of deaths in HIV-infected persons occur from HIV-associated non-AIDS (HANA) causes and people suffering from these conditions are more likely to be affected by adverse symptoms (3). An individual’s ability to identify and self-manage symptoms of HIV illness has been shown to improve patient outcomes and quality of life (4). The goal of our work was to identify the most frequently reported symptoms for PLWH with HANA conditions and integrate them into an existing web-based self-management system (VIP).

Methods

We conducted an anonymous online survey with 82 PLWH with HANA conditions to inquire about their symptom frequency and intensity in the past 30 days. Study participants were recruited from POZ.com, Craig’s List and Facebook.com (the largest online social networking site). In our survey, we asked participants to report 1) In the past 30 days, which of the following symptoms made it difficult to carry out your daily activities? 2) In the past 30 days, how much did the symptom interfere with your day to day activities?

Results

Demographics of survey participants are reported in Table 1. The top ten most frequently reported symptoms included: 1) Fatigue, 2) Muscle aches/ pains, 3) Difficulty falling asleep, 4) Anxiety, 5) Depression, 6) Neuropathy, 7) Difficulty remembering things, 8) Difficulty staying asleep, 9) Shortness of breath and 10) Decreased sex drive. The ten most intense symptoms were: 1) Fatigue, 2) Muscle aches/ pains, 3) Difficulty falling asleep, 4) Neuropathy 5) Depression, 6) Difficulty staying asleep, 7) Anxiety, 8) Difficulty remembering things, 9) Shortness of breath and 10) Decreased sex drive.

Conclusion

This study is introducing a novel research design, recruitment from online social networking sites, to study daily symptoms of PLWH, which is data that we can't get from clinical data or clinic-based surveying of this population. Findings from our survey regarding what self-care strategies they used to try to ameliorate these symptoms will be incorporated into the existing VIP system which will result in the VIP-HANA system. The VIP-HANA system will deliver these strategies to PLWH with HANA conditions.

References

Usability Evaluation of an Evidence-based Dental Patient Case Simulator

Kelsey M. Schwei, PhD¹, Kate L. Thomas, MA², Vijayakumar Thirumalai, BTech³, Chris Enstad BS³, Kim Johnson, MDH³, Andrew Schmidt, MS⁴, Olga Godlevsky, BA⁴, Neil Johnson, DDS,PhD⁵, Bill Rush, PhD⁴, Amit Acharya, BDS,MS,PhD¹,²

¹Institute for Oral and Systemic Health, Marshfield Clinic, Marshfield, WI; ²Biomedical Informatics Research Center, Marshfield Clinic, Marshfield, WI; ³HealthPartners Institute, Bloomington, MN; ⁴HealthPartners Dental Group, Bloomington, MN

Abstract

Dental Decision Simulation (DDSim) was developed by HealthPartners Institute (HPI) to aid dentists in their continued education of the latest evidence-based approaches to practice. The abstract provides a brief description of the formal usability evaluation of DDSim conducted by the study team.

Introduction

Innovative methods are needed to further educate dentists in the latest evidence-based practice. Incorporating evidence-based dentistry (EBD) into practice is currently taught in dental school curricula, but dentists face barriers to incorporating new research into their daily practice. DDSim, a novel web-based, patient case simulator for dentists to practice evidence-based care, was developed by HPI. The objective of this study was to conduct a formal usability evaluation of DDSim to discover any design issues and whether it was intuitive to dentists before planning for implementation.

Methods

A convenience sample of fifteen practicing United States dentists was recruited to participate in the study. The participants completed think aloud usability studies to discover inconsistencies or usability issues with DDSim. Usability analysts observed and recorded participants attempting to complete a treatment plan of varying complexity (hard, medium, and easy) on DDSim. Participants then completed an exit interview consisting of the System Usability Scale (SUS) survey¹ and product reaction cards from the Microsoft Desirability Toolkit.² The recordings were analyzed for themes on usability issues. SUS scores were calculated from the survey and user-selected terms from the reaction cards were counted.

Results

Participants earned an average of 23% of possible points for their respective case (hard average: 18%; medium average: 12%; easy average: 40%). Participants rated DDSim with an average SUS score of 50.8, which corresponds to a letter grade C when all SUS scores were curved like test grades. The most common terms selected from the reaction cards included complex, frustrating, usable, annoying, comprehensive, confusing, hard to use, organized, and unattractive.

Discussion

Much of the functionality of DDSim needs a redesign to ensure completion of desired actions. Participants did not earn many points while completing their cases because they only treated their given patient’s chief complaint. A majority of points in all cases were earned for treating issues the patient did not mention. Participants expressed enthusiasm for EBD, but they found the user interface of DDSim to be complex and the length of patient cases to be prohibitive to success. Further work on the interface of DDSim is needed to refine the system to make it more usable for dentists.

Acknowledgement

Research reported on this poster was supported by NIDCR of the NIH under award number RO1DE22332 and by funds from HealthPartners Institute for Education and Research.

References

An Evaluation of ‘Definite’ Anaphylaxis Drug Allergy Alert Overrides in Both Inpatient and Outpatient Settings

Diane L Seger, RPh1,2, Sarah P Slight, MPharm, PhD, PGDip1,3, Elizabeth R Silvers, BA1,2, Mary G Amato, PharmD, MPH1,4, Julie M Fiskio, BS1,2, Adrian Wong, PharmD1,4, Patrick E Beeler, MD1,5, David W Bates, MD, MSc1,6

1The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA; 2Partners HealthCare, Wellesley, MA; 3School of Medicine, Pharmacy and Health, The University of Durham, Stockton on Tees, Durham, UK; 4MCPHS University, Boston, MA; 5Research Center for Medical Informatics, University Hospital, Zurich, Switzerland; 6Harvard Medical School, Boston, MA

Abstract: Surprisingly drug allergy interaction (DAI) alerts that warn against the reaction ‘anaphylaxis’ are being overridden. We found 1,271/1,851 (68.7%) and 142/226 (62.8%) DAI anaphylaxis alerts were overridden in the inpatient and outpatient setting, respectively. A sample of these overrides where the medication being ordered was an exact match to the allergen listed (definite match) was assessed and the majority found to be appropriate.

Introduction: Clinical decision support (CDS) systems are designed to provide real-time guidance and support to providers at the point of prescribing. Drug allergy interaction (DAI) alerts are generated when a known adverse sensitivity-inducing substance is prescribed. Anaphylaxis is a serious life-threatening allergic reaction and typically a red flag for providers. However, a recent study at our institution showed that providers overrode the majority of DAI alerts including those that warned against ‘anaphylaxis.’ This study aimed to determine the reasons why providers overrode these alerts and whether these overrides were appropriate.

Methods: All DAI alerts, specifically those relating to the documented reaction of ‘anaphylaxis’, were analyzed for inpatients at Brigham and Women’s Hospital and 36 primary care practices affiliated with Partners HealthCare between January 2009 and December 2011. These interruptive alerts suggested a life-threatening reaction and gave the provider the option of either ‘cancelling’ the order or ‘overriding’ the alert. A total of 1,851 inpatient DAI alerts and 226 outpatient DAI alerts were collected. Data included the name of the drug and allergen that triggered the DAI alert, and the reasons that the provider gave for overriding the alert. A research pharmacist reviewed patient’ s charts to assess the appropriateness of these overrides that were classified as ‘definite’, i.e., where the medication being ordered was the same as the allergen listed, according to a predetermined set of criteria.

Results: Among the 1,851 anaphylaxis DAI alerts in the inpatient setting, 202 (11%) were classified as ‘definite’; in the outpatient setting, we found a total of 226 anaphylaxis DAI with 20 (9%) classified as ‘definite’. The rate of definite anaphylaxis DAI overrides was 46% (93/202) in the inpatient setting and 70% (14/20) in the outpatient setting. The appropriateness rate of the definite DAI overrides for inpatients was 83.9% (78/93) and 100% (14/14) for outpatients. The most common reason for overriding these alerts in the inpatient setting was “administered per desensitization protocol” (64) and for the outpatient setting, “patient does not have this allergy -will d/c pre-existing allergy”(8). There were 15 DAI inpatient overrides found to be inappropriate. Twelve (80%) of these overrides were intercepted by pharmacy so the patient never received the medication, while the remaining overrides were for prochlorperazine suppositories as needed, which the patients never received.

Conclusion: The rate of overrides for ‘definite’ anaphylaxis DAI alerts were found to be high but appropriate for the majority (83.9%) or all (100%) of the overrides in the inpatient and outpatient setting, respectively. Inappropriate drug allergy alerts at best can waste clinicians’ time, and the information contained in patients’ drug allergy lists should be kept updated.

References

This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ)
RuSH: a Rule-based Segmentation Tool Using Hashing for Extremely Accurate Sentence Segmentation of Clinical Text

Jianlin Shi MS MD¹, Danielle L. Mowery PhD¹, Kristina Doing-Harris PhD²,
John F. Hurdle MD PhD¹
¹University of Utah, Salt Lake City, UT; ²Westminster College, Salt Lake City, UT

Introduction
Sentence segmentation, a critical natural language processing (NLP) task, has a profound effect on downstream processing tasks, because most clinical NLP tasks are applied at the sentential level. Most sentence segmenters are developed using non-clinical text and do not segment clinical text well due to clinical text’s telegraphic style e.g., non-standard grammar, short forms, and missing punctuations. Machine learning-based segmenters trained on one clinical corpus may not perform consistently well on other corpora. On the other hand, rule-based segmenters leveraging regular expressions are rule-order-dependent, which makes the rule maintenance onerous and processing cumbersome once the rules are enriched. We present an easily adaptable, efficient, and extremely accurate sentence segmenter well suited to clinical NLP. We tested our system against three popular sentence segmenters.

Method
RuSH (Rule-based Segmentation using Hashing) leverages a nested hash table data structure to execute simultaneous rule processing, which reduces the impact of the rule-base growth on execution time and eliminates the effect of rule order on accuracy. Training and testing sets were prepared using the SemEval-2015 task 14 dataset, including 431 clinical notes of four types: discharge summaries, electrocardiogram, echocardiogram, and radiology reports. Sentence annotation was made by one annotator and reviewed by another two annotators. We randomly split the documents for training (50%) and testing (50%). The training set was used to develop RuSH rules (732 rules), as well as retrain the cTAKES segmenter model to ensure a fair comparison between the two systems. The testing set was used to evaluate the five approaches: RuSH, Stanford CoreNLP, GATE RegexSentenceSplitter, cTAKES segmenter using its default model and the retrained model. Sentence boundary detection performance was assessed using precision, recall, F1-score, and run time (overall run time and normalized run time by rule size).

Results
Our results demonstrate a notable accuracy improvement by RuSH over the other four approaches. In terms of normalized run time, RuSH demonstrates superior efficiency over GATE (100 rules) and Stanford CoreNLP (16 rules).

Table 1. Accuracy performance evaluation of RuSH against cTAKES, Stanford CoreNLP and GATE

<table>
<thead>
<tr>
<th>Segmenters</th>
<th>Sentence begin detection</th>
<th>Sentence end detection</th>
<th>Run time (ms) on testing set</th>
<th>Normalized run time (ms/rule)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
</tr>
<tr>
<td>RuSH</td>
<td>0.998</td>
<td>0.973</td>
<td>0.985</td>
<td>0.997</td>
</tr>
<tr>
<td>Stanford CoreNLP</td>
<td>0.624</td>
<td>0.786</td>
<td>0.696</td>
<td>0.597</td>
</tr>
<tr>
<td>GATE</td>
<td>0.726</td>
<td>0.774</td>
<td>0.749</td>
<td>0.704</td>
</tr>
<tr>
<td>cTAKES (default)¹</td>
<td>0.933</td>
<td>0.669</td>
<td>0.779</td>
<td>0.934</td>
</tr>
<tr>
<td>cTAKES (retrained)²</td>
<td>0.935</td>
<td>0.735</td>
<td>0.823</td>
<td>0.934</td>
</tr>
</tbody>
</table>

* Using cTAKES default segmenter model, **using model retrained with our annotated training set. ms = milliseconds

Conclusion
RuSH is an efficient, reliable, and adaptive sentence segmentation solution that can be used in an NLP pipeline either in-line or to pre-annotate clinical texts. Our purpose is to compare RuSH with other off-the-shelf solutions. Although GATE’s splitter can be optimized through customizing rules, it is difficult to handle hundreds of rules because of the aforementioned ordering issue. By adding and/or removing a small portion of rules, RuSH can quickly be adapted to other corpora. The sentence annotations will be made available upon request with SemEval data access approval.
Disease Connections at RGD

Mary Shimoyama PhD, Stanley Laulederkind PhD, Jeff De Pons BS, Marek Tutaj MS, Victoria Petri PhD, G. Thomas Hayman PhD, Shur-Jen Wang PhD, Jennifer R. Smith MS, Jyothi Thota MS, Omid Ghiasvand MS, Monika Tutaj PhD, Melinda R. Dwinell PhD
Medical College of Wisconsin, Milwaukee, Wisconsin

Abstract

An integrative software platform at the Rat Genome Database provides researchers with software tools to explore and analyze human, rat and mouse genes for disease association. Users can easily retrieve gene sets associated with single or multiple diseases and further filter by those common functions such as drug interactions or pathway participation. Resulting data subsets can be analyzed for genomic variation and additional functional features and connections.

Introduction

The Rat Genome Database (RGD) (rgd.mcw.edu) integrates comprehensive genetic and genomic data for rat, human and mouse and provides a variety of software tools to explore, analyze and visualize this data. In addition to complete gene datasets for all three organisms, variant data for Human and Rat are also available along with predictions of impact and known disease relationships for the variants. Disease-gene associations are curated from published human and rat literature using an enhanced ontology based on the Disease Ontology, MEDIC, OMIM with additions from RGD staff. RGD also provides multiple types of annotations for genes including drug/chemical-gene interactions, pathway and disease (Table). This data provides the basis for innovative software tools that allow researchers to identify subsets of genes associated with particular diseases and make connections among gene sets through commonalities in function.

Data Exploration and Analysis Tools

OLGA-Object List Generator and Analyzer, is a tool that allows users to create human, rat or mouse gene lists through easy to create complex queries using genomic regions, symbol lists or disease, pathway, drug or other functional searches. Multiple lists can be generated and a final list generated by combining lists, subtracting one from another or identifying the union of the lists. The resulting gene subset can then be downloaded or submitted to additional tools to identify variants within the genes and predicted effects or to a genome viewer to visualize location or to the Gene Annotator. The Gene Annotator provides comprehensive reports for each gene, analyzes common functional features among the genes using a distribution presentation or heatmap tool for further investigation.

Conclusion

RGD provides complete catalogues of rat, human and mouse genes and comprehensive functional annotations that can be mined, explored and analyzed using a suite of innovative tools.
Comprehensive Evaluation of Univariate and Multivariate Distributions for Machine Learning With Microbiome Data

Ali Shojaee Bakhtiari, PhD, Alexander V. Alekseyenko, PhD
Program for Human Microbiome Research, The Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC

Introduction

Human body is host to complex microbial communities collectively known as the microbiomes. The human microbiome has comparable number of cells to the number of our own cells. This additional genetic pool of the microbiome provides us with an extended metabolic and genetic capability not present in our own genome (1), and is a potential source of pathognomonic relationships in diseases. This can be exemplified in the recent linkage of the microbiome to obesity and inflammatory bowel disease.

The data for microbiome studies come from high-throughput sequencing. In particular, a marker gene sequencing assay based on 16S rRNA gene, allows for rapid and extensive identification of the microbiota. The data from this assay comes in the form of short sequences of the variants of the 16S rRNA gene present in the analyzed specimen. These sequences are processed by specialized pipelines to create standard sample-feature matrices for subsequent machine learning analysis. The minimal feature of such analyses is an operational taxonomical unit (OTU), which can be thought of as representing an underlying microbial species. The data matrix, therefore, contains counts of observed OTUs in each sample. The data are typically high-dimensional and sparse. We evaluate statistical distributions for modeling the OTU frequency vectors to recommend the most suitable ones.

Materials and Methods

We approach the problem of finding appropriate distributions for modeling the microbiome data by evaluating an extensive array of univariate and multivariate distributions to model count data of typical microbiome datasets in the model selection framework. The following univariate distributions are considered in this benchmark: Poisson, Negative binomial, Poisson Inverse Gaussian. In addition, we consider the zero inflated versions of these distributions. We use Akaike information criterion (AIC), Bayesian information criterion (BIC) and Deviance information criterion (DIC) for model comparison.

In our multivariate analyses of the data, we focus on Dirichlet-Multinomial and Beta-Liouville multinomial distributions as the two main distributions capable to account for OTU overdispersion and. The first author has shown the relative strength of the Beta-Liouville distribution compared with Dirichlet distribution in modeling of discrete count data (4).

Our benchmarking dataset consists of the data from 300 healthy participants of the Human Microbiome Project. These subjects have been sampled at 18 distinct body sites to capture the diversity of the human microbiota. A total of approximately 5,000 samples have been collected. ~45,000 individual OTUs are present in these data. We apply our methods to each OTU or the entire community in multivariate case at each body site to determine the best model(s).

Conclusion

As no single distribution has been determined to be unanimously superior, we recommend that appropriate distribution be determined empirically. We provide the set of distributions of microbiome count data utilized in this analysis as an R package: Empirical univariate and multivariate modeling for microbiome data (EUMM). The software is capable of automatically finding the optimal distribution for the provided dataset using specified model selection criterion.

References

Comparing Clinical Decision Support of a Homegrown Versus a Vendor Electronic Health Record System

Elizabeth R. Silvers, BA1,2, Diane L. Seger, RPh1,2, Adrian Wong, PharmD1,3, Mary G. Amato, PharmD, MPH1,3, Sarah P. Slight, MPharm, PhD, PGDip1,4, Patrick E. Beeler, MD1,5, Julie M. Fiskio, BS1,2, David W. Bates, MD, MSc1,6
1Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, MA; 2Clinical and Quality Analysis, Partners HealthCare, Wellesley, MA; 3MCPHS University, Boston, MA; 4School of Medicine, Pharmacy, and Health, Durham University, Stockton on Tees, United Kingdom; 5Center for Medical Informatics, University Hospital, Zurich, Switzerland; 6Harvard Medical School, Boston, MA

Introduction: Evidence suggests that CDS improves patient care, but the amount of improvement achieved is typically a fraction of what is possible. Our institution recently adopted a vendor-based system after using a homegrown EHR system for over 20 years. While the new system has medication CDS in place, it is not considered to be as user-centered as our previous homegrown EHR. We were interested in targeting providers who had used both systems and exploring their views on the design of alerts displayed in both systems. We felt that this could help identify potential areas of improvement.

Methods: We approached providers who frequently overrode medication CDS alerts (n=1,770) in the inpatient setting between January 2009 and December 2011 and interacted with the vendor system upon its deployment in May 2015. A research pharmacist trained in effective academic-detailing techniques conducted these sessions, which were tailored to each provider’s particular overrides. Graphical materials including performance level data were presented to each provider and used as the basis for a two-way discussion. A robust and complete analysis using a qualitative analytic tool (NVivo 11 for Windows) of the data was carried out, and key concepts related to views on alert functionality and specific prescribing behavior were identified.

Results: We conducted a total of 35 sessions, which included a diverse group of providers of varying professional backgrounds including physicians, physician assistants, nurse practitioners, and pharmacists. Providers felt several aspects of the CDS in the vendor system were less effective than the homegrown EHR system including the timing and tiering of alerts. Most providers felt the lack of tiering of CDS in the vendor system increased their alert fatigue in comparison with the homegrown EHR. They also found the presentation of all alerts at the time of signing an order, as opposed to at the time of ordering, to be overwhelming. Many agreed that drug-drug interactions (DDI) should be considered high priority in the vendor system as in the homegrown system. CDS of renal-based alerts outside of labeling renal insufficiency was rarely offered, whereas the homegrown EHR offered dose and frequency adjustments and appropriate alternatives, which caused providers to often refer to outside resources to make informed decisions. Distinguishing between true drug allergies and intolerances in the new EHR was an improvement, though many were unaware of this functionality. Providers suggested that the new vendor system incorporate alert tiering and draw attention to serious reactions and contraindications through the use of hard stops.

Conclusions: A number of insights surrounding alert functionality in both EHR systems were identified. Regarding DDIs and renal-based alerts, providers felt the homegrown system provided more useful information. Most were unaware that the vendor system distinguished between true drug allergies and intolerances: therefore attention should be drawn to this feature in provider training. Providers felt they overrode more alerts and experienced greater alert fatigue using the new system, though ascertaining these alert override rates is beyond the scope of the current study. These findings will form the basis for several recommendations to senior management on how to improve the vendor system. By incorporating a more robust CDS system, the vendor EHR could more effectively help clinicians with their decision making and deliver safer patient care.

References:

This study was funded by grant #U19HS021094 from the Agency for Healthcare Research and Quality (AHRQ)
Introduction

Metadata-driven information systems provide end users with the ability to manage information models that otherwise require developers to write code. Metadata-driven design also allows for auto-creation of forms and data display interfaces as well as the use of common data elements. Use of research Common Data Elements is increasingly being encouraged by funding agencies (1). CDEs have the potential to increase the harmonization of data collected in clinical trial management systems. National Institute of Dental and Craniofacial Research (NIDCR) Office of the Clinical Director (OCD) oversees many clinical trials and needs to manage the lifecycle of these trials effectively. This requires a flexible data model that also supports data collection in a standardized way.

Material and methods

To enable effective management of clinical trials data by the NIDCR OCD we created a web service that allows the end users to model their data using a metadata-driven approach. This tool is integrated with National Institutes of Health National Library of Medicine’s Common Data Elements (CDE) repository so the users can incorporate CDEs into their data models with ease. The web service is created using open source technologies, including Linux Ubuntu, Apache, PostgreSQL, Ruby on Rails web framework (2), Bootstrap, and JQuery. The system allows complex type specific validation rules to be defined for each CDE. For example a text field may have a regex-based validation where as an integer field may have a range defined as the validation rule. Fields may have more than one validation rule set which allows the value domain for each user defined field or CDE to be precisely defined.

Results and Discussion

The web service is now being used at the NIDCR OCD to manage 34 clinical trials. Since its inception, data managers and research nurses have used the service to create data models. A total of 17,415 CDEs from (https://cde.nlm.nih.gov/home) have been integrated into user defined fields and can be used while designing new tables and forms. A central user defined field repository contains all published fields from individual projects as well as imported common data elements from public repositories such as the NLM.

Conclusion

A metadata-driven system allows seamless integration of common data elements into data models designed by end users and minimizes the need for programmer involvement in information system design.

References


This work was supported by the Intramural Research Program of the NIDCR, NIH.
Developing an Electronic Health Record-Based Cohort of Patients with Inflammatory Bowel Diseases for Observational Patient-Centered Outcomes Research

Siddharth Singh, MD, MS1; Paulina Paul, MS1; Chun-Nan Hsu, PhD1; Lucila Ohno-Machado, MD, PhD1,2

1University of California San Diego, La Jolla, CA and 2San Diego Veteran’s Administration Health System, San Diego, CA

Introduction

Observational research in inflammatory bowel diseases (IBD) has traditionally been limited by small sample size and event rates (in single-center studies) or lack of detailed phenotype analysis (in administrative health claims databases). We propose to overcome these barriers by creating a large contemporary electronic health record (EHR)-based cohort of patients with IBD seen across any of the five University of California (UC) medical campuses, which combines structured (and codified) data with NLP-derived phenotype data, utilizing the platform developed for the patient-centered SCAlable National Network for Effectiveness Research (pSCANNER),1 a clinical data research network that is part of PCORnet. A combination of structured and narrative data has been shown to have superior performance to identify and characterize patients in “phenotyping” activities for numerous diseases.

Balancing sensitivity and specificity

The cohort for this study will be identified using a sequential two-step process as shown in the Figure. In the first step, we will screen all records using ICD-9 codes for Crohn’s disease (CD) (ICD 9 555.x) and ulcerative colitis (ICD 9 556.x), to identify a large cohort of patients with a potential diagnosis of IBD. To assess feasibility we used the UC Research eXchange Data Explorer (powered by SHRINE) to estimate a total of 23,296 potential patients with IBD, of whom 5,726 (24.6%) appear to have been treated with anti-TNF agents. From a cohort derived from pSCANNER, we will develop an algorithm, based on structured data elements, to more accurately identify patients with CD and UC. Structured data elements include: number of ICD-9 codes for CD and/or UC, diagnostic codes for competing diagnosis (irritable bowel syndrome, diverticulitis), codes for inpatient hospitalization, gastroenterologist visit or endoscopic procedure, diagnostic codes for CD or UC-related complications (intestinal fistula, stricture, perianal fistula or abscess), procedural codes for abdominal or perianal surgery (using Current Procedural Technology – CPT codes), laboratory values for elevated inflammatory markers (such as C-reactive protein), and prescription of IBD-related medications in the EMR prescription program. However, there are limitations related to use of only structured data elements for IBD diagnosis. Hence, to improve upon our approach to identifying patients with IBD, we will analyze narrative text (within clinical notes, endoscopy reports, etc.), using NLP techniques to augment the identification process of IBD patients. The accuracy of the combined models at various specificity levels will be calculated using a hierarchical generalized linear model, and the overall prediction performance evaluated based on the area under the receiver operating characteristic curve (AUC), using a training set of randomly selected 500 patients with IBD. This will be initially validated in an independent test set of 100 patients for whom manual chart review will be completed.

In summary, our approach combining structured and narrative data, using NLP algorithms, will help to identify and characterize a large, EHR-based contemporary cohort of IBD patients across the UC system for use in observational comparative effectiveness and patient-centered outcomes research.

References


This work is funded by ACG Junior Faculty Development Award, PCORI CDRN-1306-04819 and NIH CTSA grant UL1TR001442.
Pilot Evaluation of Usability and Acceptability of a Mobile Application for Congestive Heart Failure Readmission Reduction: The HealthPal Project

Minal A. Singhee, MS¹, Rema Padman, PhD¹, Sriram Iyengar, PhD², Daniel Gartner, PhD³, Robert Monte, MBA⁴, Ashley Ketterer, MHA⁴, Brianna Scott, MS⁴, Angelo Baiocchi, RPh⁴
¹Carnegie Mellon University, Pittsburgh, PA, USA; ²Texas A & M University, College Station, TX, USA; ³Cardiff University, Cardiff, United Kingdom; ⁴Veterans Engineering Resource Center, VA Pittsburgh Health System, Pittsburgh, PA, USA

Introduction

Congestive Heart Failure (CHF) is a major chronic condition characterized by the inability of the heart to pump sufficient blood to the body. CHF readmission is a critical contributor to the burgeoning healthcare cost¹. Drawing on social cognitive theory and persuasive computing technology, we have developed HealthPal, a customizable, multimedia-rich post-discharge education, follow-up and monitoring application for CHF that uses an innovative technology platform called guideVue for creating and publishing apps on mobile devices². The aim is to help CHF patients understand their condition better, improve knowledge and enable self-efficacy in post-discharge self-health management to potentially reduce the risk of readmissions. This study is a preliminary evaluation of the factors associated with the usability, acceptability and workflow aspects of HealthPal.

Methodology

We developed a fixed protocol for the exploration of the app via interviews with all the stakeholders. This included the Think Aloud Protocol³, a quiz to understand the content and ease of use of the application and a survey to obtain feedback from users. Multiple analyses were performed to understand the usability, acceptability and workflow features of the app, including qualitative analysis of the aggregated free text responses in the Think Aloud Protocol and content quiz, a quantitative analysis of the survey via summary statistics and Key Log Sequence Analysis of app navigation. Figure 1 shows a complete exploration flow to test user’s navigation patterns and completeness of their app exploration to analyze workflow. We encoded the five major sections of the application as follows: a. Exercise, b. Blood Pressure, c. Cholesterol, d. Risk, e. Resources

Results

In-person interview results from 30 stakeholders including patients, physicians, nurses, case managers and social workers indicate that the app is very informative and easy to use. Table 1 summarizes the average scores from patients’ feedback tested on a scale of 1 to 5. Users suggested improvements in the layout of the app to make it more accessible, and recommendations for desirable interactivity functions in a future version.

Conclusion

Patients and clinicians reported positive experiences with the interactive mobile technology. Future extensions of HealthPal will incorporate additional interactive features and voice recognition capabilities to facilitate personalized patient-provider interactions.

References


Table 1. Patient Feedback Summary

<table>
<thead>
<tr>
<th>Average Scores</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Usefulness</td>
<td>3.7</td>
</tr>
<tr>
<td>Ease of Use</td>
<td>3.4</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>3.8</td>
</tr>
<tr>
<td>Learning</td>
<td>3.7</td>
</tr>
<tr>
<td>User Interface</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Figure 1. Complete Navigation Flow

Figure 2. HealthPal Screenshot
Tracking Risk of Acute Mental Status Change in VA Hospitals

Stacey L Slager, BA, MS1,2, Bryan Gibson PT, PhD, 1,2 Teresa P Taft, BEd1,2, Nancy Staggers, RN, PhD1,2, Charlene R Weir, RN, PhD

1Biomedical Informatics, University of Utah, Salt Lake City UT; 2 George E. Whalen Veterans Affairs Medical Center, Salt Lake City, UT

Introduction

Nurses spend more time with the patient than any other care provider and are responsible for the bulk of patient care. Thus, they are most likely to notice when the patient is experiencing an acute mental status change (AMSC), referring to changes in brain function, such as confusion, loss of alertness, loss of orientation, defects in judgement or thought, unusual or strange behavior, poor regulation of emotions and other functions. Acute mental status changes can have a significant impact on patient care in terms of symptom management, patient safety, and staffing; therefore monitoring changes in mental status, initiating preventative actions and mitigating harm requires effective coordination between clinical roles. Formal assessments for AMSC and tracking that information over time is lacking. Well-defined data from patient assessment for AMSC can aid development of decision support in a learning health organization.

Methods

Semi structured interviews were conducted with 58 Nurse Managers (NM) of acute medicine floors across regional areas of Veteran’s Affairs hospitals in the US. Our purpose was to characterize nurses’ medication management information needs, documentation and communication processes associated with identifying and monitoring AMSC in inpatient settings. We asked questions about admission assessment and documentation of AMSC; identifying patients at risk by formal assessment or otherwise; interventions for patients at risk and associated policies regarding risk; and documentation and planning at the ward/institution level. The interviews were analyzed with descriptive statistics.

Findings

When asked how mental status was assessed upon admission, 49(84.5%) nurse managers said it was part of the nursing admission process. Only 7(12.0%) reported a formal assessment other than orientation, such as with the Confusion Assessment Method (CAM). Most NMs (n=54, 93.1%) documented orientation in the nursing admission template. Risk for delirium was identified by: Admission diagnosis, prior diagnosis, prior delirium (n=24, 41.4%) with clinical judgement (n=10, 17.2%) and flagging of behavioral history (n=6, 10.3%) being the next most common method approaches. While the assessment of patients for AMSC is somewhat informal, it is a key aspect to making nursing assignments (n=46, 79.3%). The majority of NMs did not know if AMSC was tracked on an institutional or ward level, but they do track related measures (e.g. sitters, falls, etc.) (24, 41.4%). Table 1 highlights additional results.

Table 1. selected results from Nurse Manager interview

| Used a specific tool for delirium assessment | 6(10.3%) | Said there was a policy for contacting physicians if patient at risk for delirium 8(13.8%) |
| No formal risk assessment for delirium | 44(75.9%) | Interventions for patients at risk of delirium: 16 unique interventions mentioned |
| Documented delirium risk in CPRS 8(13.8%) | | |
| Notifies physician when change in cognition, behavior | 36(62.1%), 21(36.2%) | Do charge nurses have separate report of AMSC? Yes 22(37.9%) No 29(50%) |

Discussion

Mental status assessment is limited largely to orientation and no policies are in place that uses risk measurement to drive increased monitoring. Identifying risk appears to not be a formal process.

References

A Literature Review of The Approaches Used to Train Qualified Prescribers to Use Computerized Provider Order Entry Systems

Clare Brown, MPharm¹,², Katie Reygate, MPharm PGDip³, Ann Sleg, MPharm⁴,⁵, Jamie J. Coleman, MD MA⁵, Sarah K. Pontefract⁵, MPharm PGDip⁵, David W. Bates MD, MSc⁶,⁷,⁸, Andrew K. Husband, PhD¹, Neil Watson MBA MSc², Sarah P. Slight, PhD¹,²,⁶

¹School of Medicine, Pharmacy and Health, Durham University, U.K. ²Newcastle upon Tyne hospitals NHS Foundation Trust, U.K.; ³Health Education KSS Pharmacy, U.K.; ⁴eHealth Research Group, Centre for population health sciences, University of Edinburgh, U.K.; ⁵College of Medical and Dental Sciences, University of Birmingham, Birmingham, U.K.; ⁶The Center for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Women’s Hospital, Boston, MA, USA; ⁷Harvard Medical School, Boston, MA, USA; ⁸Harvard School of Public Health, Boston, MA, USA.

Abstract: Studies have suggested that insufficient training is associated with suboptimal use of Computerized Provider Order Entry (CPOE) systems. We conducted a literature review to describe the approaches used to train qualified prescribers on these systems, whether online methods in particular were used and to identify whether they included the challenges of using these systems. A number of different approaches were used to train prescribers, including demonstrations, practical exercises and ward-based training. Educating prescribers about the pitfalls and challenges of CPOE systems was rarely discussed.

Introduction: Computerized Provider Order Entry (CPOE) systems have been associated with potential benefits over paper-based systems, including improved patient safety. A key element of the implementation and on-going use of these systems is ensuring that users are, and remain, sufficiently trained to use them effectively. Studies have suggested that insufficient training is associated with suboptimal use.¹ We sought to describe the approaches used to train qualified prescribers to use CPOE systems, whether online methods in particular were used, and to identify whether they covered the pitfalls and challenges of using these systems.

Methods: We performed a literature review, using a systematic approach. Three large databases: Cumulative Index Nursing and Allied Health Literature (CINAHL), Embase and Medline were searched. We did not restrict the timeframe for these searches in any of the respective databases. Articles that explored the training of qualified prescribers (medical and non-medical practitioners) on systems in both inpatient or outpatient settings were included. Reference lists of included articles were also examined. A narrative synthesis of all eligible studies was undertaken.

Results: Our database search returned 1,155 publications. After reviewing the titles, abstracts and full text, a total of seven were included. One further article was identified from the reference list search. A variety of training methods were used, such as: demonstrations; expertise-specific exercises aimed at addressing commonly encountered prescribing errors (e.g. prescribing an antibiotic at non-standard times); and assessments, which allowed users to monitor their own progress. Super-users played a valuable role in providing ward-level support and reducing the need for external training, with potential cost-savings. Web-based demonstrations were used in one study. A team at the University of Victoria in Canada developed an online portal that housed a range of simulated versions of electronic health record systems, where individuals could prescribe for fictitious patients. Educating prescribers about the challenges and pitfalls of CPOE was rarely discussed and a comparison between methods was lacking.

Conclusion: We identified few articles related to the training of prescribers on the use of CPOE systems. Approaches used include demonstrations, practical exercises and ward-based training. Using a combination of approaches was considered important. Online learning may facilitate the training of a large number of users, offering them the opportunity to practice and become familiar with the system at a convenient time and place.

References

Needs Assessment: A Subscription-based Laboratory Notification System

Benjamin H. Slovis MD¹,², Hojjat Salmasian MD MPH PhD¹,³, Gilad Kuperman, MD PhD¹,³, David K. Vawdrey PhD¹,³

1. Department of Biomedical Informatics, Columbia University, NY, NY, 2. Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, NY, NY, 3. Quality and Patient Safety, New York-Presbyterian Hospital, NY, NY

Introduction
The delayed review of laboratory results is potentially harmful. Established processes (e.g. phone-calls) provide notification of critical laboratory values, however physician awareness of non-critical and normal results also affects clinical decision-making. Previously described notification systems have demonstrated improved physician responses to critical results, yet few studies have documented subscription-based notifications. Allowing providers control of notifications for laboratory results may reduce delays in patient care. We propose a tool to provide near-real-time notifications of laboratory results through text-page and email, via subscription at the time of order-entry. This system will allow clinicians to subscribe to individual tests and receive notifications when results are available. To inform the function of the system, we solicited input from resident physicians.

Methods
A survey instrument created with Qualtrics software (Qualtrics, Provo, UT) was distributed via email to all house-staff in the medicine and pediatrics residencies to determine the perceived usefulness of a subscription-based laboratory notification system. The survey included questions regarding the frequency of lab order and review, perceived barriers to result review and physician opinions on our proposal. Responses were analyzed using descriptive statistics and qualitative responses were coded to generalized themes.

Results
A total of 63 residents responded to the survey, accounting for 29% of the 216 eligible house-staff. The majority (87%: 55/63) of the respondents reported repeatedly reviewing labs throughout the day, with 60% of these repeatedly checking labs more than 5 times per day. (Table 1). Sixty-two percent (39/63) of respondents reported frustration from repeatedly checking lab results, while 63% (40/63) reported that repeatedly checking results caused delays in clinical care. When asked about our proposed tool, 59% (37/63) of respondent reported they would find it useful. When asked about preferred methods of notification, 83% (52/63) of respondents reported that they would carry a pager, with only 41% (26/63) stating they would prefer SMS and 16% (10/63) preferring email. The majority of residents (53%) who responded to questions regarding how notifications would affect patient care implied it would save time in their clinical workflow. (Table 2)

Table 1. How often do clinicians have to check the EHR to see if a test they are waiting for has resulted?

<table>
<thead>
<tr>
<th>Repeatedly Checking Results?</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5 times per day</td>
<td>6%</td>
</tr>
<tr>
<td>6-10 times per day</td>
<td>35%</td>
</tr>
<tr>
<td>&gt;10 times per day</td>
<td>31%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2. In what ways would clinicians expect notification to improve clinical care?

<table>
<thead>
<tr>
<th>Improvement</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save Time</td>
<td>53%</td>
</tr>
<tr>
<td>Improve Prioritization</td>
<td>44%</td>
</tr>
<tr>
<td>Notification of Critical Result</td>
<td>3%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

Conclusion
Our respondents were generally dissatisfied by the current state of result review in the EHR, and were accepting of our proposed subscription-based laboratory result notification system. The next steps in our research are to deploy the notification system and evaluated whether its use reduces time between available results and physician awareness.
eMeasure Certification: Validating Logic for Electronic Clinical Quality Measure (eCQM) Reporting
Anne Marie Smith, RN, BSN, MSHSA1, Ben Hamlin, MPH1
1National Committee for Quality Assurance (NCQA), Washington, DC

Abstract
NCQA developed an electronic process to verify a software package has the electronic clinical quality measures (eCQMs) correctly programmed and the resulting Quality Reporting Document Architecture (QRDA) Category 1 file is correctly generated. NCQA is certifying systems that produce these QRDA 1 files to provide data for quality reporting.

Introduction
The eMeasure certification process is designed to validate measure-specific software code for extracting the correct electronic health record (EHR) data as defined by CMS eCQM specifications. NCQA’s system offers more robust testing than other systems, because it tests many more cases per measure and includes testing for negative events. These measure-specific tests use a black-box testing model to create a comprehensive set of measure scenarios based on the eMeasure specifications and known measure-specific programming errors. The goal of certification is to pre-approve the integrity of the software code to ensure its adherence to measure specifications and its ability to accurately produce comparable results.

Methods
NCQA creates unique sets of sample data or "test decks," for each eCQM, developed from randomly generated patient-level test data. Each test deck contains between 800-1000 individual patients designed specifically to test each measure’s initial population, denominator, exceptions, exclusion and numerator inclusion logic. The test data also includes negative testing, or testing for data not applicable to the measure. Test decks are generated from flow diagrams that allow NCQA to mark the portions of the measure to which each sample patient applies. A vendor seeking NCQA certification process a test deck through their measure code and returns results for each patient in the test data. The vendor results for each patient are then compared to the expected results to determine if the vendor software code computes the measure correctly. NCQA also validates the vendor’s ability to report the correct results in the QRDA Categories I and III standard document format. Multiple test decks are created for each eMeasure in order to correct errors identified from each pass of the deck through the vendor code. Each measure is tested individually using Continuity of Care Documents (CCDs). Each vendor that completes certification receives a final certification report and seal which indicates the measures certified.

Results
NCQA has completed this process with 3 vendors and beta tested with a 4th. Every vendor had to make adjustments to their measure calculation logic to match NCQA’s expected results. Common errors included age calculation errors, missing results of lab tests, vendor defined exclusions not in measure specification, misspelled elements in the QRDA file, and codes not included in the value set but present in calculation. Vendors were most challenged in the certification process by measures containing an index event for denominator eligibility or those containing specific timing linked to a denominator event in order to determine numerator compliance.

Conclusion
Vendors who achieve certification will improve the precision and long-term data comparability while also having pre-approved data, which may reduce costly chart reviews.
Human Factors Analysis of Look-Alike/Sound-Alike Medication Errors

Katrina H. Smith1, Aaron Z. Hettinger, MD MS1,2, Allan Fong, MS1, Maryann E. Amirshahi, PharmD MD MPH2,3, Grace M. Tran, MS1, Erica L. Savage1

1MedStar Health National Center for Human Factors in Healthcare, Washington, DC; 2Georgetown University School of Medicine, Washington, DC; 3MedStar Washington Hospital Center, Washington, DC

Introduction
Look-Alike/Sound-Alike (LA/SA) medication errors have been estimated to account for approximately 25% of all medication errors1. In order to recognize these errors, the Institute for Safe Medication Practices (ISMP) maintains a list of the most common LA/SA name pairs that have the potential to be confused2. Given the great potential for these events to result in patient harm, a human factors analysis of the frequency and sources of LA/SA errors was undertaken.

Methods
In an effort to develop potential solutions for preventing LA/SA medication incidents, a patient safety event database was reviewed to understand the contributing factors leading to medication cases involving LA/SA drug names. The database, containing roughly 90,000 incident reports over a 3-year span, consists of frontline staff self-reported near misses or harm events in an academic healthcare system. The free text narrative field of the reports was searched using all 399 unique keyword pairs from the ISMP list of confused drug names2. An additional search was performed using approximate string matching to account for spelling mistakes3. Each case narrative was reviewed independently by two subject matter experts (AZH, MEA) and categorized according to one of four phases in the medication use system: 1) documentation, 2) ordering, 3) dispensing, or 4) administration of medication. Error severity level was also determined based on standard toxicology outcome definitions, taking into account both anticipated and actual patient outcomes. Cases of disagreement were discussed until consensus was reached.

Results
A total of 131 initial events were retrieved, of which 82 incidents met inclusion criteria consistent with a LA/SA error. These 82 incidents represent 12 different entities within a single health system. Five of these 131 events (4%) produced an initial disagreement between reviewers, and only one required discussion between reviewers before reaching a consensus. The most common LA/SA medication pairs in the analysis were oxycontin-oxycodone (20%), hydroxyzine-hydralazine (12%), ceftriaxone-cefazolin (11%), prednisone-prednisolone (9%), sulfasalazine-sulfadiazine (7%), hydromorphone-morphine (6%), oxycodone-hydrocodone (6%), dopamine-dobutamine (5%), and risperidone-ropinirole (5%). These pairs accounted for 80% of identified errors. Analysis by date indicated a mean of 2.48 events per month (standard deviation 1.84) over the 33-month period, and a range of 0-7 events per month. The most common reporters were pharmacists (45%), registered nurses (38%), and administration/management staff (9%). The most common errors phase was dispensing (50%), followed by ordering (26%), administration (21%), and documentation (4%). Severity level categorized 46% of events as moderate, 41% as mild, 7% as severe, and 1% as critical. Near misses accounted for 61% of events, while 39% of reported errors reached the patient. Of those events that reached the patient, 9% resulted in temporary harm, while the other 91% resulted in no harm.

Discussion and Conclusion
Initial analysis of the LA/SA medication errors confirms that further exploration is warranted, particularly within the dispensing phase. Future research will review the common themes and discuss the vulnerabilities in medication practices with the health system’s pharmacy team. This will provide a deeper understanding of workflow practices and additional insights into trends within specific facilities. Potential human factors solutions will be formulated with the pharmacy team to develop system-based improvements that will prevent the occurrence of LA/SA errors.

References
1. Lambert B, Lin S, Chang K, Gandhi S. Similarity as a risk factor in drug-name confusion errors: the lookalike (orthographic) and sound-alike (phonetic) model. Med Care. 1999;37(12);1214-1225.
Categorization of Patient Health Goals Elicited During Home Care Admission

Paulina Sockolow DrPH MBA MS¹, Kavita Radhakrishnan RN PhD MSEE², Edgar Chou MD, MS¹
¹Drexel University, Philadelphia, PA; ²The University of Texas, Austin, TX

Introduction

Home care agencies (HCA) provide patients with skilled nursing care and ancillary services at home; however, HCA patient rehospitalization rates exceeded 30% for numerous health conditions.¹ In home care, a program of patient self-management goal elicitation with behavioral change decreased hospital readmission and improved health outcome.² Our objective was to categorize elicited patient health goals and identify “clinically informative” goals.

Methods

The research team with a HCA partner examined patient goals that admitting clinicians documented in the point-of-care electronic health record (EHR) during a 5-month pilot project in 2015. Patient goals were available in a text string in a de-identified Excel file that the HCA extracted from their EHR. To develop a coding scheme, a researcher (PS) conducted content analysis on patient goal data: 1-assigned themes to the first 100 patient goals; 2-grouped themes into codes; and 3-specified code categories. A home care nurse (KR) reviewed the coding scheme that PS used to assign a goal code to every 10th patient. Records without a patient goal were tabulated. PS added new codes that emerged to the coding scheme that KR reviewed. Subsequently, KR and the physician researcher (EC) reviewed the coding scheme independently to identify codes that were informative to their disciplines (clinically important).

Results

Of the 1,763 patient records, 18% had no recorded patient goal. After content analysis of 122 records, the coding scheme totaled 20 codes among 3 categories: 12 Health Management (HM); 6 Activities of Daily Living (ADL); and 2 Quality of Life (QoL). In the sample of records with patient goals, there were 1 to 4 goals documented in each record, for a total corpus of 253 goals. Most goals were phrased in clinician vernacular (e.g., “increased ambulation”) and 6 were in a patient’s voice (e.g., “to be able to walk again”). Codes identified as clinically important to both disciplines were equally distributed among the ADL category (6 codes; e.g., ambulation, safety) and the HM category (7 codes; e.g., manage disease process, nutritional status) with no QoL codes selected. There were 5 clinically important codes that also occurred most frequently: safety/falls (ADL, 18%); ambulation (ADL, 9%); ADL activities (ADL, 9%); manage disease process (HM, 9%); knowledge of disease process (HM, 10%).

Discussion

The absence of the patient’s voice and less than universal recording of home care patients’ goals indicated differential clinician documentation of elicited patient goals. Consistent communication of the intent and operationalization of patient goal elicitation may address differential documentation. In addition, clinician training may be advisable to have clinicians understand why they are asking patients about their goals.¹ Findings also suggest that the most frequently occurring codes were codes identified as clinically important for home care nurses and primary care physicians. These findings indicate a shared perspective about the importance of specific clinical information in the treatment of home care patients; however, a Norwegian study found differences in perspectives.³

Conclusions

Research is needed to identify the most effective approach to operationalize patient goal elicitation; clinically important goals using a larger group of clinicians; and optimal dissemination of this information in patient care. Useful research would also be to identify associations between elicited patient goals, nursing interventions, and outcomes.

References

Health Reform in Minnesota: An Analysis of Complementary Initiatives Implementing Electronic Health Record Technology and Care Coordination

Karen L. Soderberg, MS¹, Sripriya Rajamani, MBBS, PhD, MPH², Douglas R. Wholey, PhD², Martin LaVenture, PhD, MPH, FACMI¹

¹MN Department of Health, St. Paul, MN; ²University of MN, Minneapolis, MN

Introduction

In 2008 the Minnesota Legislature enacted sweeping health reform intended to improve the affordability, access and quality of care to improve the health of Minnesotans¹. Among other initiatives, the law established certification of patient-centered health care homes (HCH). A year prior, another groundbreaking effort was enacted when the Minnesota Legislature passed the interoperable electronic health record (EHR) mandate requiring all health providers in Minnesota to implement interoperable EHR systems by 2015². We examine synergies between these two programs by studying the association between adoption and use of EHRs in primary care clinics and HCH certification.

Findings

By end of 2014 nearly all of Minnesota’s HCH eligible clinics have implemented EHR systems, including 99.7% of certified HCH and 98.9% of not certified clinics. Overall, HCH clinics were better utilizers of EHR systems for the metrics examined in this study: use of clinical decision support functionalities, utilization of patient disease registries, use of EHR data to support quality improvement efforts, electronic exchange of summary care records, and availability of patient portals (Figure 1). Our results differ from other studies of EHR implementation in Patient Centered Medical Homes, which found more variable adoption of care coordination tools than our study found³. A potential source of this difference is the convergence of complementary health reform initiatives in Minnesota has provided opportunity to show distinct differences in the utilization of EHRs among clinics that have committed to EHR adoption and HCH certification. Certified HCHs have more effectively implemented clinical decision support tools that can result in improved care, improved patient safety, and lower costs. Certified HCHs also better utilized electronic patient registries, allowing them to better track patients with chronic conditions in order to improve care, observe outcomes, and monitor progress toward care plan goals.

Conclusion

The rapid uptake in EHR technology, combined with health reform efforts that focus on accountability and care coordination, pose challenges and opportunities for clinical care providers. Challenges include managing changing workflows, training staff, and understanding system capabilities. Opportunities are real-time access to current patient information, use of automated decision support tools, and secure sharing of information with the patient’s entire care team. These findings demonstrate the value that EHRs can bring to achieving coordinated patient care.

References

Asthma Ascertainment NLP System Portability across Institutions

Sunghwan Sohn, PhD1, Yanshan Wang, PhD1, Chung-Il Wi, MD2, Elizabeth A. Krusemark2, Euijung Ryu, PhD1, Mir H. Ali, MD3, Young J. Juhn, MD2, Hongfang Liu, PhD1
1Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN; 2Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN; 3Department of Pediatrics, Sanford Children’s Hospital, Sioux Falls, SD

Introduction
Clinical NLP systems have proven to be successful in various tasks. However, the performance often varies across institutions. Whenever a NLP system developed in one corpus is applied to another corpus, the questions arise: “how similar are the two corpora?” and if two corpora differ “how does the difference affect the NLP system portability?” In this study we examined similarities of asthma birth cohorts between Mayo Clinic and Sanford Children’s Hospital (SCH) in terms of entire corpus at word-level and asthma-related concepts. Then, we further explored the NLP system portability for asthma ascertainment.

Materials and Methods
The study cohorts were selected randomly from Mayo and SCH patients (n=298 for each) and matched with age (median = 2.3 years). The cosine similarities between Mayo Clinic and SCH corpora were compared using vector representation defined as follows: 1) Corpus: the entire corpus of each institution was compared as a whole using tf-idf (term frequency-inverse document frequency) and latent topic-based vector. The tf-idf for the term t is defined by summation of tf(t)idf(t) for all documents in the corpus divided by total number of documents in the corpus (N), such as ∑tf(t)idf(t)/N. In the latent topic-based vector representation, the topic zk for the corpus C is defined as p(zk|C) = ∑d∈C p(zk|d)/N; 2) Asthma-related concepts: the concepts were extracted and compared. A vector representation of asthma-related concepts for each corpus was created using the definition of cf-idf (concept frequency-inverse document frequency), where we use the concept frequency in the corpus instead of the term frequency.

Results
Table 1 contains the similarities between Mayo Clinic and SCH corpora. Figure 1 shows the distribution of asthma-related concepts between Mayo Clinic and SCH. The NLP asthma ascertainment system was originally developed on the Mayo’s sick child cohort and tested on the Mayo’s late-preterm cohort with an F-score of 0.762. The current NLP system has been refined on the late-preterm cohort. However, it has not been elevated yet on the held-out test set. The out of box NLP system was applied to the SCH cohort and produced an F-score of 0.813.

Discussion
The entire corpus similarity was merely mediocre showing the different nature between the two corpora in the word level. However, the NLP system for asthma ascertainment is largely dependent on asthma-related concepts and therefore we believe that concept-wise similarities are more important as an estimate of the NLP system portability than the word-level similarity. The direct comparison of the NLP system performance has not been performed due to the lack of Mayo Clinic’s gold standard annotation. However, we believe that the comparison to the previous study shows a reasonable insight about the performance in SCH. The Mayo Clinic and SCH corpus were relatively homogeneous in concepts that show a good potential for the NLP system portability. However, a proper adjustment will be necessary to deal with intrinsic corpus heterogeneity in order to produce the desirable performance of asthma ascertainment.

Table 1. Similarity of Mayo and SCH.
<table>
<thead>
<tr>
<th></th>
<th>tf-idf</th>
<th>topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus</td>
<td>0.669</td>
<td>0.944</td>
</tr>
<tr>
<td>concepts</td>
<td>0.971</td>
<td>NA</td>
</tr>
</tbody>
</table>

Figure 1. Distribution of asthma-related concepts (y axis is proportion)
Automated Extraction of Disease Activity Component Measures From Electronic Medical Records Using Natural Language Processing

Brett R. South, MS, PhD1,2, Shobhit Mehrotra1,2, Chris Leng, MS, Chao-Chin Lu, PhD1,2, Brian C. Sauer, PhD1, Grant W. Cannon, MD1
1VA Salt Lake City VA Health Care System, Salt Lake City, Utah, 2University of Utah, Department of Biomedical Informatics, Salt Lake City, Utah

Abstract Chart Review is often used to obtain patient data only found in clinical texts such as patient outcomes or measures of disease activity. These efforts have high cost and may be less efficient than using natural language processing (NLP). Automating retrieval of disease activity score measures would be a significant benefit for pharmaco-epidemiologic research on rheumatoid arthritis (RA). We evaluate a rules-based NLP approach to extract disease activity scores from VA clinical documents.

Introduction The Veterans Affairs rheumatoid arthritis (VARA) registry is a multi-center observational study of RA patients collecting data from 12 VHA medical centers including 28 joint disease activity score (DAS28) using manual chart review. This method of retrospective data collection is inefficient and automated NLP extraction of DAS28 scores that are entered by providers during a patient encounter using a semi-structured note template would be a great advance. NLP systems have shown promise in reducing labor-intensive chart review tasks such as quality assessment1, risk assessment2, cohort generation3, and for extracting RA disease activity4.

Methods We constructed a development set of 472 clinical documents for VA Salt Lake City VARA patients (n=207) between January 1, 2015 and September 30, 2015. This set was used to train a rules-based NLP system to extract three DAS28 core measures: tender joint count (TJC), swollen joint count (SJC), and patient global assessment (PtGA). A validation set was constructed using 484 clinical documents from the same site for VARA patients (n=214) seen in clinic between April 1, 2014 and December 31, 2014. We evaluate system performance metrics (accuracy and precision) comparing NLP outputs against clinical gold standard data obtained from the VARA registry by manual review.

Results Accuracy and precision calculated using the validation set and treating the VARA registry as the clinical gold standard for each of the three DAS28 core measures was: TJC (93%, 100%), SJC (92.1%, 100%), and PtGA (92.1%, 99.8%). NLP failures were all explained by situations where incorrect note templates were used or where the template was modified.

Conclusion The development of an NLP system to electronically retrieve disease activity measures would be a significant benefit for pharmaco-epidemiologic research on rheumatoid arthritis (RA). Our process was successful at retrieving DAS28 components from clinical documents and can be implemented in an automated way to reduce chart review burden. Future directions will apply these methods to other VARA sites and automate population of NLP retrieved data into the VARA registry.

References

Contrasting Autism Subgroup Genotypes Using Frequent Pattern Mining

Matt Spencer¹, Chi-Ren Shyu, PhD¹,²,³,⁴

¹Informatics Institute, ²Department of Computer Science, ³School of Medicine, ⁴Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO

The etiology of autism spectrum disorder (ASD) remains poorly understood, imposing an obstacle to the diagnosis and treatment of people with the disorder. Genome-wide association studies (GWAS) have identified isolated gene candidates that are only minimally associated with the disorder¹, so recent studies emphasize the effects of interacting single nucleotide polymorphisms (SNPs)². Most studies search for genetic causes of autism as a whole, rarely considering that autism may be an assortment of many disorders with similar phenotypes. If this is the case, then sets of SNPs may contribute to specific subdivisions of the disorder. Here we apply the idea of interacting SNPs to identify genetic differences between potential subpopulations of autism.

We analyzed the Simons Foundation Autism Research Initiative (SFARI) dataset comprising over 2600 families, a larger population than most autism studies. SNP microarray genotype data and phenotype classification data (demographics, behavioral phenotypes, etc.) are available describing the proband, both parents, and siblings for each family. Thus, every affected individual is contrasted by unaffected individuals with many of the same SNPs. We exploit this to focus our analysis to the combinations of multiple SNPs (multi-SNPs) unique to the individual with autism in each family.

Frequent pattern mining is a useful method for identifying combinations of genotypes that may be related to notable phenotypes, especially since genotype data introduces far too many variables for a regression analysis. We previously implemented a frequent pattern mining package in Apache Spark³, allowing us to utilize a distributed computing environment to address the computational challenge of testing combinations of many genetic loci. This software identified multi-SNPs that are frequently observed in individuals with autism.

We developed a workflow that allows us to highlight multi-SNPs that show a significant difference in prevalence between two subpopulations. After stratifying a population according to phenotype classifications, we calculate the occurrence of multi-SNPs in each subpopulation using the frequent pattern mining package described above; these are subsequently filtered when the multi-SNP is not unique to the proband within each family. The populations are then contrasted by comparing the frequency of occurrence of each multi-SNP in either group.

Our preliminary study contrasts probands with and without substantial physical abnormalities. We focus on variations in genes that have been studied in the context of autism, a total of 740 genes derived from the AutDB database⁴. We identified 55 multi-SNPs that have at least a five-fold difference in prevalence in these populations, providing candidate genes for further investigation into the underlying genetic distinction between the groups. This method is being utilized in ongoing research to compare other subpopulations of the Simons dataset.

We aim to use this method to uncover distinctions between subpopulations of autistic children and provide candidate genes that are associated with specific phenotypes for further investigation. This may lead to a better understanding of the mechanisms that produce individual phenotypes and may even provide evidence that a subpopulation should be considered a separate disorder, allowing for more specific diagnosis and treatment of each group. Although the etiology of autism is complex and will not be explained through any single effort, an emphasis on multi-SNPs allows this method to suggest contribution from genes that would otherwise be overlooked by conventional GWAS methods.

References

Smartphone-platforms Apps and Medical Disciplines serving Mobile-Health

Basile Spyropoulos\textsuperscript{1,2} PhD, Maria Bania\textsuperscript{2} BSc, Irini Mathe\textsuperscript{2} BSc, Stavros Miloulis\textsuperscript{2} BSc, Georg Tziokas\textsuperscript{2} BSc, Kostas Proestos\textsuperscript{2} BSc, Iliana Gerakou\textsuperscript{2} BSc, Manolis Kapernaros\textsuperscript{2} BSc
\textsuperscript{1}Technological Educational Institute (TEI) of Athens, GR, \textsuperscript{2}University of Athens, GR

Abstract

A “3-dimensional project” is presented addressing, firstly, cardinal medical-managerial m-Health aspects, secondly, evaluating currently available Smartphone-OS/platforms and finally attempting to design Apps to be employed in important m-Health Medical Specialties. It is based on the outcomes of a novel, innovative and collective final-examination approach, in the MSc.-Course “Medical Information Technology and Telemedicine”, for the seven Students of the common Graduate-program “Information Technologies in Medicine and Biology” of the University of Athens and the TEI of Athens, based on extended reviews of Patent Applications and other Papers.

Introduction

Mobile-health is a term used in Medical Practice, supported by mobile devices, collecting clinical data, healthcare-information, allowing for real-time monitoring of vital signs etc. The Internet of Things (IoT) holds a great potential to meet yet unmet needs of Healthcare, however, IoT technologies demand further R & D. In this paper, we attempt to predict the potential of the combination of these emerging approaches, as reflected mainly on Patent Documents.

Methods and Resources

The most important aspects of m-Health have been approached as a “triplet” of components including (Table 1):

\begin{itemize}
  \item Medical-managerial m-Health aspects as Health-vigilance, Safety, Market-trends, Providers’ Liability etc.
  \item The advancement of the OS/platforms developed, as reflected on relevant published Patent-Documents.
  \item An attempt to design Apps, based on similar principles, to support m-Health related Medical Specialties.
\end{itemize}

328 m-Health relevant Patent Applications (PA) have been retrieved and evaluated, by employing the esp\@cenet search-engine of the European Patent Organization (EPO) and a detailed mapping has been created (2000-2016).

Results

The group has focused on the most urgent or promising issues and aspects of m-Health as presented in Table 1.

Table 1. An overview of the three major components of the collective project presented.

<table>
<thead>
<tr>
<th>Mobile-Health aspects addressed</th>
<th>OS/Platform</th>
<th>Medical Specialties supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Health-vigilance and Training</td>
<td>Android</td>
<td>Emergency Medicine (at home, en route, in working place etc.)</td>
</tr>
<tr>
<td>Effective, lasting equitable and fair access to health-services</td>
<td>BSD</td>
<td>Tele-monitoring and Tele-nursing services (ECG, ENG, PO\textsubscript{2}, O\textsubscript{2} Saturation, Oximetry etc.)</td>
</tr>
<tr>
<td>Patient-data Security and Interoperability</td>
<td>iOS</td>
<td>Tele-medication and electronic order entry</td>
</tr>
<tr>
<td>The m-Health Hardware-market</td>
<td>Linux</td>
<td>Tele-recovery and tele-rehabilitation</td>
</tr>
<tr>
<td>The m-Health apps-market and the Providers’ Liability</td>
<td>OS X</td>
<td>Tele-Psychiatry, Tele-Neurology and Tele-Psychological Support</td>
</tr>
<tr>
<td>Patient Safety and medical-technical standards</td>
<td>Windows Phone</td>
<td>Tele-care of Trauma and minor Telesurgery</td>
</tr>
<tr>
<td>Reimbursement methods and Liability of Health-care Personnel</td>
<td>IBM z/OS</td>
<td>Tele-imaging and mobile Medical Imaging systems Tele-ENT, Tele-Ophthalmology etc.</td>
</tr>
</tbody>
</table>

Conclusion

The reviewed publications and PA-documents contain hints that m-Health Devices and Apps are gradually being related, to the emerging Internet of Things and it seems that they will play an important role in the near future.

1604
Analyzing Similarities in Handoff Communication Content between Residents and Nurses

Vignesh Srinivasana, Thomas Kannampallilb, Trevor Cohena, Joanna Abrahama
a University of Illinois at Chicago, Chicago, IL, United States; b University of Texas School of Biomedical Informatics, Houston, TX, United States

Abstract
Past empirical handoff research has emphasized the importance of developing shared understanding among various clinicians (physicians, nurses) for streamlining the handoff process. We compared the degree of content overlap between the content of communication of handoff between residents and nurses using two methods: first, using a standardized Jaccard’s coefficient, and second, using human raters. We found significant association between the two methods ($r=0.54$) suggesting that Jaccard’s coefficient is a reliable measure for analyzing clinical content.

Introduction
Handoffs are primarily a communication activity that involve a transfer of patient information, responsibility, and authority between clinicians. A vast majority of handoff tools are designed to promote discipline-specific communication. While recent studies have emphasized the importance of building an interdisciplinary ontology for handoff communication, we have identified two gaps in the current handoff research: (1) a lack of understanding of the potential overlap in the content of communication between disciplines; and (2) lack of an analytical method that can identify this content overlap between disciplines. However, no reliable methods, besides human coding, exists for establishing the overlap between content. In this paper, we compare the human coded transcripts of overlap with Jaccard’s coefficient to evaluate its appropriateness as a metric for investigating content overlap.

Method
The study was conducted in an Internal Medicine unit of a large academic hospital in Midwestern United States that has over 13,000 hospital encounters per year. Handoffs between outgoing and incoming residents/nurses occurred every day at shift changes. Resident and nurse handoffs of 10 random patients were shadowed and audio-recorded. Transcripts of residents ($n=10$) and nurses ($n=10$) were segmented along the clinical content categories based on a previously developed handoff content coding framework (that includes identifying information, code status, allergies, disposition, medical history, active problems, assessment of active problems, pending labs/procedures, medications/treatments, tasks/to do, family)². Human rating was provided by two clinical experts (1 resident and 1 nurse): For each pair of transcripts (resident and nurse), the rater indicated how similar they perceived the transcripts were on a five-point scale with 1 indicating “no similarity” and 5 indicating “high similarity” (this was done for each content category). These pairs of documents were blinded and randomized to all coders. Jaccard’s coefficient has been used to effectively measure the similarity between words with high stability. Next, for each pair of documents, we computed the Jaccard’s coefficient. The ratings obtained from human coding and the Jaccard’s method were then compared using Pearson’s correlation.

Results & Discussion
We obtained a moderately high degree of correlation between the two rating methods ($r=0.54$). While we had a relatively small set of documents, the Jaccard’s coefficient achieves a good correlation with human judgement. The results clearly show that the Jaccard’s coefficient, which is widely used for comparing the similarity and diversity of sample datasets, is a reliable measure to evaluate similarity in communication content between disciplines.

References
Most Common Allergen Terms in the DoD and VA
Developing a Baseline Set of Allergen Terms for the C-CDA Using RxNorm "Ingredient" Level Abstraction

Karl J. Stiller PharmD, David A. Parker MD, Brett A. Marquard, BS, Margaret A. Marshburn, RN, Sharon L. Moore, BS, Steven F. Kator MD, Norman E. Stone III MD

DoD/VA Interagency Program Office, Rosslyn, VA

Objective
Determine the most common allergen terms recorded in the Departments of Defense (DoD) and Veterans Affairs (VA) electronic health records, with a goal of improving semantic interoperability for drug-allergen clinical decision support, and establishing a baseline set of terms for the Consolidated Clinical Document Architecture (C-CDA).

Methods
The enterprise data stores of the DoD and VA were queried for all allergen terms and corresponding frequency of use. Medication terms were mapped to RxNorm Concept Unique Identifier (RxCUI); Medication classes were mapped to the National Drug File - Reference Terminology (NDF-RT). Medications were then abstracted to the RxCUI "Ingredient" level. The DoD and VA lists were merged to a single list using RxCUI and NDF-RT as the common keys. The list was rank-ordered by combined frequency of use and the 95% usage list was calculated.

Results
601,666 recorded allergies to medications or medication classes were analyzed. The DoD data store contained a dictionary of 103,862 terms, of which 4,339 terms were used at least once. 446 terms represented 95% of all recorded allergens. The VA data store contained 2,298 terms, all of which were used at least once. 305 terms represented 95% of all recorded allergens. Combined, the Departments’ lists contained 498 unique medication and medication class allergen terms. The top 3 allergen terms were Penicillin, Lisinopril, and Sulfonamides (medication class), representing 14%, 8%, and 7% of all recorded allergens, respectively.

Conclusion
Relatively few allergen terms comprise the majority of recorded allergens in the DoD and VA. Agreement on this limited set of common allergen terms could simplify achievement of semantic interoperability in this important domain of health data.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Allergen Term</th>
<th>Frequency of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Penicillin</td>
<td>14%</td>
</tr>
<tr>
<td>2</td>
<td>Lisinopril</td>
<td>8%</td>
</tr>
<tr>
<td>3</td>
<td>Sulfonamides</td>
<td>7%</td>
</tr>
<tr>
<td>4</td>
<td>Codeine</td>
<td>5%</td>
</tr>
<tr>
<td>5</td>
<td>Atorvastatin</td>
<td>3%</td>
</tr>
<tr>
<td>6</td>
<td>Simvastatin</td>
<td>3%</td>
</tr>
<tr>
<td>7</td>
<td>Morphine</td>
<td>3%</td>
</tr>
<tr>
<td>8</td>
<td>Gabapentin</td>
<td>2%</td>
</tr>
<tr>
<td>9</td>
<td>Amoxicillin</td>
<td>2%</td>
</tr>
<tr>
<td>10</td>
<td>Tramadol</td>
<td>2%</td>
</tr>
</tbody>
</table>
Is More eHealth System Use Better for Cancer Patients and Family Caregivers? A Literature Review

Wu-Chen Su, M.S., Ming-Yuan Chih, Ph.D.
College of Health Sciences, University of Kentucky, Lexington, Kentucky, USA

Abstract
We aim to review eHealth literature for cancer patients and their family caregivers, and summarize differential eHealth system use effects. 2595 articles were searched, and 12 that met the exclusion/inclusion criteria were reviewed. The review demonstrated that more system use were associated with improved health outcomes. However, further studies are needed to provide stronger evidence and application of differential eHealth use effects.

Introduction
EHealth has been shown to support cancer patients and their family caregivers. However, such interventions do not always lead to predictable outcomes due to different system use. Increasingly, researchers went beyond reporting randomized trial results by examining the impact of system use on the effects of eHealth interventions. We aim to provide a review of these studies and summarize the reported differential system use effects.

Methods
In February 2016, the authors searched PubMed using these broadly defined eHealth terms: “((Internet OR web* OR smartphone* OR "mobile app") AND (therapy OR “self-help” OR intervention)) OR eHealth OR mHealth OR telehealth OR telemedicine OR [Mesh] "telehealth" } AND Cancer” and found 2595 research articles. After excluding articles that met the criteria: 1) non-full text, 2) non-English, 3) not eHealth interventions for cancer patients or caregivers, and 4) not reporting system use and users’ physical, psychological and social health outcomes, 12 papers were included. Types of cancer, intervention, study period, correlations between users’ characteristics and system use, and correlations between users’ health outcomes and system use were summarized.

Results
Of these 12 articles, 11 were published after 2010; nine conducted in the U.S. Seven articles focused on breast cancer, while the rest covered mixed cancers (including prostate, colorectal, and lung). Cancer patients and survivors were the main target populations. Only one study included family caregivers. The main studied eHealth interventions included online social support groups (6), decision aids (3), expert consultation (1) and the combination of them (2). All interventions were web-based. Most studies (8) lasted for more than four months. All studies allowed participants to use eHealth interventions freely without required or prescribed amount of use. Nevertheless, one study made the eHealth intervention available to both randomized groups but provided the study group participants structured approaches to encourage more system use. Four studies reported the correlations between users’ characteristics (e.g., psychological status at pretest, demographics and treatment history) and the level of system use. The cancer knowledge, mental illness status and perceived social support at pre-test stage are important factors to predict the level of system use. For instance, the high-use group participants were often those who reported poor psychosocial status at pretest than those in the low-use group. These 12 articles reported a total of 30 unique system use measures, including 13 basic system use measures (e.g., logins, time spent, and page views) and 17 advanced system engagement measures (e.g., posts and votes). Overall, more system use has been associated with improved physical outcomes (e.g., length of survival), psychological outcomes (e.g., depression and anxiety), self-efficacy, knowledge, behavioral outcomes (e.g., intention and attitude) and social outcomes (e.g., bonding and social support). However, two studies reported the relations between the level of system use and health related outcomes were not statistically significant, and two studies found negative impact of system use: 1) more page views were found to be related to less favorable attitude toward cancer screening tests; and 2) those who were instructed to post more emotional, informational, and companionate support messages in an online support group reported higher depression and anxiety afterwards. Detailed findings were provided here (http://tinyurl.com/hq3sebg).

Conclusion
In conclusion, several eHealth use measures and their differential effects on health outcomes of cancer patients and caregivers were found and summarized. Further studies should focus on providing stronger evidence (ideally via randomized trial design) of such a dose (system use)-response (health outcomes) relationship. Additionally, the knowledge of differential system use effects may provide the basis for the design of a more effective eHealth intervention that facilitates beneficial system use behavior of cancer patients and their family caregivers.

References

Katherine A. Sward, PhD RN¹, Alex Bui, PhD², Jose-Luis Ambite, PhD³, Michael Dellarco⁴
¹University of Utah, Salt Lake City, UT; ²University of California, Los Angeles, CA; ³University of Southern California, Los Angeles, CA; ⁴National Institute of Biomedical Imaging and Bioengineering (NIBIB), Washington, DC

Abstract
The Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) program was launched in 2015 to develop sensor-based, integrated health monitoring systems for measuring environmental, physiological, and behavioral factors in epidemiological studies of pediatric asthma. PRISMS will develop an open platform and associated suite of tools to enable a comprehensive spectrum of biomedical research into chronic diseases in pediatric populations. We describe the PRISMS program goals, operations, and organization, and informatics challenges.

Introduction
The NIH launched a cross-institute initiative to examine the effects of the environment on children’s health, including the Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) Program¹. As part of its mission NIBIB conducts research and development of new bioengineering techniques and devices to improve the detection, treatment, and prevention of disease, including technologies for disease detection and assessment of health status.

Problem addressed/purpose
Pediatric asthma is complex, with interactions between environmental, physiological, and behavioral factors. In PRISMS, researchers will develop a variety of novel sensors, sensor-based integrated health monitoring systems, and an interoperable data management platform to make environmental and health data available to epidemiologists and clinical researchers. The integrated effort will offer new insights into non-invasive health monitoring and pediatric asthma research. PRISMS is organized to address several informatics challenges.

Program organization
Sensor Development Centers develop new sensors or redesign existing sensors to monitor environmental stressors and physiological parameters correlated with pediatric asthma. Informatics Platform Centers manage secure data acquisition and processing, integration of sensor data with other data sources, and secure transmittal to the Data and Software Coordination and Integration Center, which verifies data quality, harmonize data into a consistent model, and supports analysis. Key challenges to be addressed in the PRISMS Program include developing robust measurement devices to collect environmental and behavioral data. Devices must be feasible and acceptable for study participants and investigators to use. Informatics platforms must be flexible, interoperable repositories that embrace diverse approaches to environmental and behavioral measurement; calculate interim measures such as multiple asthma severity scores; and must be generalizable to other types of environmental epidemiology studies. The platforms must accommodate mobile health (mHealth) approaches and must be scalable for large multi-center studies. Other challenges include managing uncertainty and diverse quality inherent in sensor data, the need to integrate data collected at multiple levels of granularity and timing, and information models and standards that may not adequately encompass the detail needed for sensor-intensive research.

Conclusion
Developing a sensor-based, integrated health monitoring system for measuring environmental effects on health offers challenges that require informatics innovations, methods and tools. Pediatric asthma research is likely to benefit from the comprehensive spectrum of biomedical informatics efforts developed through the PRISMS program.

References
Never-Ending Medical Learning

Douglas A. Talbert, PhD; William Eberle, PhD; Mei Liu, PhD

1Tennessee Tech. Univ., Cookeville, TN; 2Univ. of Kansas Med. Center, Kansas City, KS

Introduction and Background

Effective collection and application of healthcare experience data presents an ongoing challenge to the quality and cost of healthcare delivery in the United States and has contributed the Institute of Medicine’s (IOM) development of a vision for a continuously learning health care system to “promote and enable continuous and real-time improvement in both the effectiveness and efficiency of care.” In support of this vision, we have designed and implemented an early stage prototype that seeks to apply recent developments in never-ending machine learning to healthcare. We call this prototype the Never-Ending Medical Operative (NEMO).

The never-ending machine learning paradigm seeks to construct computer programs that, like people, learn many different types of knowledge from years of diverse experience, make use of previously learned knowledge to learn new types of knowledge, and use skills such as self-reflection to sustain continuous learning and avoid performance plateaus.

NEMO was inspired and informed by CMU’s Never-Ending Language Learner (NELL). NELL was designed to explore the application of never-ending learning to the task of reading the web by pursuing just two goals: (1) read the web to improve the quantity and quality of its knowledgebase and (2) learn to read better. Like people, NELL employs multiple strategies to achieve these goals in an ongoing, continuous manner.

Prototype Design and Implementation

Our NEMO prototype implements a step toward never-ending learning, in which semi-autonomous, adaptive learning is directed by domain experts and user feedback regarding system output. Figure 1 shows NEMO’s high-level architecture. The Client Web App is the primary user interface and is where domain experts direct and give feedback on NEMO’s learning. The AI Controller is responsible for the learning as well as adapting in response to user feedback. The Data Loader pulls a subset of records from the i2b2-based KUMC clinical data repository and stores them in the Data Mart. The Data Mart holds a subset of patient data to be used by the learners, the questions posed by the domain experts will pose, and the user feedback.

Next Steps

This prototype is the first step toward NEMO’s long-term vision of being a collection of intelligent agents that cooperatively contribute to an ever-learning computational healthcare support environment. The evaluation of this prototype will focus primarily on its ability to successfully implement semi-autonomous, adaptive learning in a medical domain. Subsequently, the scope of NEMO will expand to include work on how NEMO selects which knowledge to pursue and how NEMO can enrich its knowledge and influence health and healthcare through interactions with patients and providers.

References

3. http://rtw.ml.cmu.edu/rt

1609
Fitbit™ Fitness Tracking Ease of Use and Utility: Preliminary Findings for Potential Use in Clinical Care

David P. Taylor, PhD1; Nathan C. Hulse, PhD1,2; Chaitanya K. Mynam, MS1; Bhanu Iyer1; Matthew H. Ebert, MS1; Jason P. Gagner, MBA1; Peter J. Haug, MD1,2
1Homer Warner Center, Intermountain Healthcare, Salt Lake City, UT; 2University of Utah, Salt Lake City, UT

Introduction

Sedentary behavior is an increasing problem and a contributor to obesity, chronic disease, and rising healthcare costs. Fitness tracking may help individuals more accurately monitor and increase physical activity levels. Tracking other lifestyle aspects such as diet and sleep may also provide benefits. As part of a diabetes prevention program at Intermountain Healthcare we built tools, utilizing Fitbit’s API, to help patients share their Fitbit data and provide care team access to summarized views of the data. We are reporting preliminary findings on the ease of use and utility of the Fitbit system in a sample of test subjects prior to its use with actual patients.

Methods

We produced iterative designs for the web-based tools using input from clinicians. After implementation, 6 medical informaticists used Fitbit devices to test the system. We interviewed participants at 9 months after data were first collected with follow-ups at 22 months. The main goals of the interviews were to assess the ease of use of the Fitbit system and determine the utility that the participants derived from it. We accessed summarized data on physical activity, sleep, diet and weight through the data review tool to assess how often and how consistently test subjects were tracking.

Results

The web-based patient enrollment and fitness data review tools were successfully implemented in a production environment. The enrollment tool was available through the Intermountain patient portal, My Health. An automated nightly call retrieved any new or changed Fitbit data for participants through the API. The fitness data review tool displayed summarized views of the data for authorized care team members. Over 22 months, test subjects tracked steps 73% of possible days with a daily average of 10,383 steps. Four of six test subjects continued to use the system regularly at 22 months. Sleep and diet tracking features were not well used (19% of nights for sleep and 1% of days for diet tracking). No participants reported body weight using Fitbit more than once or twice.

Discussion

Overall, test subjects regarded the Fitbit system as worthwhile in motivating them and/or reinforcing fitness habits, particularly for step-oriented activities that required no manual effort to track besides remembering to wear the device. Any tracking activity that requires effort, such as manually entering data, is much harder to maintain. It is obvious that individuals have different sources of motivation to engage in and track physical activity. With regard to tracking, some individuals may have more internal sources of motivation while others prefer competition, status, or other social benefits. Others may respond to monetary rewards or prizes that are more tangible. In this study the social features (e.g., comparisons with peers) were motivating to some but not others. We anticipate that a variety of methods—reminders, incentives, and frequent communication between patients and care team members may be necessary to help sustain patients’ efforts in fitness tracking. Although medical informaticists may not be representative of the general population, what we learned is still valuable. If informaticists are not willing to use a certain set of features consistently, it seems unlikely a less tech-savvy population would either.

Conclusion

The ability to acquire data from systems such as Fitbit provides interesting opportunities to integrate fitness tracking into clinical care. We have demonstrated it is possible for individuals to maintain consistent tracking habits over a period of months, or even years. However, there is still a question whether a fitness tracking device on its own, without any additional incentives or affiliated coaching, will demonstrate a significant improvement in fitness outcomes in most individuals. Until systems become even more automated and easy to use, individuals may need additional incentives and outside support to form and maintain tracking habits to reach health goals. Additionally, as more patient-generated data become available for clinical use, there are interesting questions about the utility of these data, how to summarize them appropriately, which members of the care team should review and act on them, and where these activities fit in clinical workflow.
Relieving Operational Burdens of Participants in Inter-organizational Care Coordination

Sidney N. Thornton, PhD¹-², Shan He, PhD³, Iona Thraen, PhD²³, Deepthi Rajeev, PhD⁴
¹Intermountain Healthcare, Salt Lake City, UT; ²University of Utah, Salt Lake City, UT; ³Utah Department of Health, Salt Lake City, UT; ⁴HealthInsight, Salt Lake City, UT

Introduction

Provider organizations within the Utah geographical region have demonstrated a willingness to participate in inter-organizational care coordination projects that improve the health of patients for whom they share treatment. Care coordination by definition spans multiple patient encounters and potentially multiple provider organizations using disparate EHR systems and connectivity networks. Many pilot care coordination projects are underway from provider, payer, and public health sectors¹. The current investment to participate in a care coordination project is substantial and requires both technology and personnel. The question under investigation by the Utah community is what aspects of care coordination can be pooled centrally. Policy constraints in the community favor a federated model but precedence exists for centralized repositories².

Methods

Active projects of care coordination were collected from stakeholders including initiatives funded privately and through public grants. The convenient sample of fourteen use cases of care coordination represented clinical, public health, patient engagement, and payment reform sponsorship. The technical interoperability requirements along with the workflow implications of data exchange and integration were analyzed for common functional components. Optimal integration strategies were evaluated based on existing and proposed data sharing policies and agreements, and were evaluated based on scalability for future care coordination scenarios. Gaps in policy were identified and escalated to the community governance committees for assistance.

Results

Potential economies of shared care coordination utilities and services are identified for the following electronic maintenance and discovery transactions: patient identity proofing and identity verification; patient designated proxy and legal care representation; patient’s current care team by role and active status; care team providers’ delivery preference and connectivity details; and care coordination logic (trigger recognition, service orchestration, and system and patient-registered device notification). The shared utilities and services will be hosted centrally and available for consumption by all participating stakeholders. Maturity and readiness of the proposed services are variable with most experience in the patient identity space. Standard, interoperable definitions and models for care teams are not widely implemented particularly when integrated with patient portals. Significant policy and technology gaps exist for sharing patient proxy information.

Discussion

A trade-off exists between scalable workflow automation and the operational burdens associated with care coordination. Larger organizations with more advanced clinical exchange functionality will require scalable automation to ensure replicable and standardized data behavior, while smaller facilities may tolerate manual processes in the beginning. Some independent facilities will find economies and operational relief through the state designated clinical exchange authority that will have full access to the envisioned shared services and utilities for care coordination. Limitations of the process include the breadth of the convenient sample of scenarios and a reasonable evaluation methodology for national scalability. Near term refinements of policy are being organized.

References

Development of Test Topics for Cohort Identification

Tamara Timmons, M.D.1, Stephen Wu, Ph.D. 1, William Hersh, M.D. 1
Oregon Health and Science University, Portland, Oregon

Introduction

Patient cohorts are a fundamental component for many types of biomedical research. Our long-term goal is to develop sound information retrieval (IR) methods for patient-level cohort identification from clinical text. To evaluate these methods while protecting patient privacy, we have begun creating a secure, intra-institutional resource using the Cranfield approach for evaluation of a cohort retrieval task: creating a test collection with three components: a collection of documents, a set of topics, and relevance assessments. In this abstract, we focus on the subject of topics, describing our development process and the considerations and challenges specific to their use in patient-level clinical IR.

Test topic development

The topic development process had 3 main steps: 1. definition of topic format, 2. selection of representative sources, and 3. abstraction of source documentation into topic contents.

Topics are provided in 3 different formats for possible queries: 1) summary statement; 2) brief summary and clinical – a shorter summary statement plus a mock clinical case incorporating a patient and scenario which typify the topic criteria; 3) brief summary plus structured data – a summary statement plus criteria list. This format is similar to topics in some TREC ad hoc collections, and both reflect varying user approaches1.

We aimed to have 50 topics in our set, an adequate set size to achieve evaluation measure stability2. For use as topics, we considered 60 cohorts from sources representative of different use cases for cohort identification such as clinical study recruitment or quality measure reporting. The resulting topics are drawn from 5 sources: OHSU’s research data warehouse (RDW) (29 topics), Mayo Clinic’s RDW (2 topics), Phenotype KnowledgeBase (PheKB) (7 topics), Rochester Epidemiology Project (REP) (9 topics), and the National Quality Forum (NQF) (12 topics). Similar topics were merged (1 OHSU/REP, 2 OHSU/PheKB), and 1 was excluded (non-patient cohort).

Topics vary in eligibility criteria, but a few contained extensive lists of criteria, or multiple combinations of different criteria satisfying overall eligibility. For practical reasons, these topics required simplification by removing criteria less integral to the topic’s objective.

Development of such topics requires consideration of many factors that influence the results. In developing test topics for this project, such factors included: variation in topic criteria complexity, maintaining applicability to multiple use cases, and creating multiple representations of each topic.

References

Real-Time Patient Satisfaction Surveys via an Emergency Department Portal

Marc Tobias, MD¹, Julie Westover, BS¹, Daria Ferro, MD¹, Joseph J Zorc, MD, MSCE¹

¹The Children’s Hospital of Philadelphia, Philadelphia, PA, USA

Background: The standard methods for measuring patient satisfaction in healthcare settings are limited as they are typically completed well after the encounter and have low response rates. Methods to capture patient feedback in real-time would allow clinical teams to respond and improve care prior to discharge.

Aim Statement: We assessed use of a website to measure patient satisfaction during an Emergency Department (ED) visit. We compared two methods for survey administration: 1. Invitation by research associates (RAs) to participate in providing feedback, and 2. Self-guided completion by families using their own device via a web portal (BYOD).

Methods: We created a web portal (http://ed.chop.edu) that patients and their families can access while receiving their ED care (Figure 1). A link to a RedCap™ patient satisfaction survey was offered through this web portal as a method to measure satisfaction. RAs also randomly selected patient rooms to administer an identical survey. The survey measured overall satisfaction and specific elements of care on a 1 (worst) to 7 (best) Likert scale.

Outcomes: During the evaluation period starting January 2016, RAs administered 324 complete surveys and there were 41 complete BYOD surveys. Ongoing data collection is in progress for both groups. The median overall satisfaction for the RA surveys was 7 out of 7 (IQR 6-7) with a mean of 6.6 (95% CI 6.5-6.7); 13.3% (95% CI 9.8-17.5%) of these surveys provided free-text comments that were deemed actionable. For the pilot BYOD surveys, the overall satisfaction ratings median was 7 out of 7 (IQR 5-7) with a mean of 5.9 (95% CI 5.4-6.4) and 51.2% (95% CI 35.1-67.1%) surveys contained actionable free-text comments (Figure 2).

Conclusion: Preliminary results suggest that real-time satisfaction surveys administered by RAs measure a high degree of satisfaction with relatively few actionable comments. Pilot surveys completed spontaneously by families using their own devices showed lower satisfaction with more actionable free-text comments. Further data will be available for presentation at the meeting.

Figure 1. Screenshot of web portal from a mobile device

Figure 2. “How would you rate this ER overall?”
Operationalizing Best Practice for First Follow-up Visit after Discharge from a Well-Baby Nursery
Michael Totzke, Catherine Staes PhD, Julie Shakib D.O. MPH,
Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City

Background
The American Academy of Pediatrics (AAP) Bright Futures guidelines for the timing of the first well-child visit (WCV) after discharge from a Well Baby Nursery (WBN) recommend that the visit occur within 48 hours of discharge for newborns with a WBN length of stay of <48 hours, and within 3 to 5 days for those with a WBN length of stay of >48 hours.¹ The purpose of these early WCVs is to detect conditions that may cause readmission in the first weeks after birth. Recommendations that the first outpatient visit occur within a few days after discharge differ from the previous standard practice of the first WCV at 2 to 4 weeks of age.² A recent large retrospective study by Shakib et al. demonstrated that completing the first WCV in compliance with the AAP guidelines was associated with a 15% reduction in the rate of readmissions compared with a later first WCV.³ We hypothesized that adding a “recommender” module or application within the electronic medical record (EHR), such as a best practice alert (BPA), may help to improve compliance with the AAP guidelines and prevent newborn hospital readmissions. The logic that would determine best practice could be integrated into an EHR as a set of rules that could function as a type of “recommender” to a clinician in the nursery.

Methods

Process modeling
To understand the processes and the logic of a clinician’s workflow involved after the birth of a baby, as the mother and baby are being observed and preparing for discharge, I interviewed Dr. Julie Shakib, medical director of the Well Baby and Intermediate Nurseries at the University of Utah Healthcare. Dr. Shakib described each step in the process flow that a pediatrician goes through when documenting the first follow-up appointment using the Epic EHR system.

Operationalizing business logic
The logic was documented which describes all the data needed to be captured in the workflow involved in the processes of setting a first follow-up appointment. The logic was placed in sequential order and then reviewed and refined with Dr. Shakib by adding clinical criteria that would influence the timing of the first appointment. We compiled a list of clinical criteria and diagnoses, and their associated ICD-9 codes, that might appear in the problem list portion of the EHR and which might change the scheduled time of discharge as well as the scheduled first follow-up appointment.

Discussion
A problem with the current process of documenting the first follow-up appointment is evident when discharge planning begins. In the Epic EHR, when a pediatrician initially wants to document a patient’s discharge information, they first have to enter the name of the outpatient pediatrician before they can even see the discharge form. An improvement to this process would be to present the discharge information, including the anticipated first follow-up date and time, to the pediatrician the first time they click on the discharge form. Getting the information with less effort and in less time will help the process move more smoothly.

There are several foreseeable scenarios that might hinder the ability to make the appointment within the recommended time frame. For example, the recommended date and time might fall on a weekend. Saturdays are sometimes possible appointment days, but pediatric clinics are generally not open for appointments on Sundays. In these cases, the attending pediatrician could use their judgement to decide if the baby could be seen on the following Monday, or would need to be seen earlier. This decision could be documented as free text in the discharge note or, better yet, as a structured comment in a new data field added to the discharge tab so the rationale is more easily computable.

References

Introduction
Patients presenting to the emergency department (ED) with life-threatening aortic dissections (AAD) can be challenging to diagnose. Though the time-sensitive condition is rare, these patients are at high risk for death; therefore delays in diagnosis must be avoided. AAD symptoms can be vague and may mimic other conditions such as acute coronary syndrome or abdominal pain. Our team examined a population-based ED database to identify presenting symptoms of patients diagnosed with AAD to enhance current clinical decision rules. Upon review of the data, we noted that the patient records included AAD in all 11 diagnosis (dx) positions (from 1-primary, to 2-secondary…to position 11). Prior to analyzing symptom data, we had to differentiate patients with life-threatening AAD from those with a history of AAD. The purpose of this study was to develop a case definition for ED visits attributable to AAD by determining the accuracy of including visits with AAD dx in secondary or tertiary positions.

Methods
We obtained ED visit data for 2008-12 from a statewide ED database which accounted for 99.5% of all visits to North Carolina (NC) EDs. Included were visits with an ICD-9-CM code for AAD (441.00-441.03) in any dx position. We then grouped the visits into 11 strata based on the dx position containing the AAD code. We also conducted a manual chart review of a stratified random sample (AAD in dx positions 1-11) of 2008-12 visits to an academic medical center ED in NC. A trained research nurse manually reviewed the full EHR record for each patient to determine whether the visit was attributable to AAD or a history of AAD.

Results
2841 (0.01%) of the 22.4 million ED visits in NC had an AAD dx, distributed across all 11 dx positions (Table 1).

<table>
<thead>
<tr>
<th>Dx position</th>
<th>N (%)</th>
<th>Dx position</th>
<th>N (%)</th>
<th>Dx position</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>978   (34%)</td>
<td>5</td>
<td>168   (6%)</td>
<td>9</td>
<td>43   (2%)</td>
</tr>
<tr>
<td>2</td>
<td>725   (26%)</td>
<td>6</td>
<td>104   (4%)</td>
<td>10</td>
<td>25   (1%)</td>
</tr>
<tr>
<td>3</td>
<td>440   (15%)</td>
<td>7</td>
<td>70    (2%)</td>
<td>11</td>
<td>12   (&lt;1%)</td>
</tr>
<tr>
<td>4</td>
<td>223   (8%)</td>
<td>8</td>
<td>53    (2%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

31 charts with dx 441.00-441.03 were reviewed. 16 (52%) of the visits, with AAD in positions 1-3, were for life-threatening AAD (Table 2). The remaining 15 had unrelated primary dx, plus a history of aortic dissection in the past.

<table>
<thead>
<tr>
<th>Dx position</th>
<th># visits for AAD/total # records reviewed (%)</th>
<th>Dx position</th>
<th># visits for AAD/total # records reviewed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9/9 (100%)</td>
<td>4</td>
<td>0/1 (0%)</td>
</tr>
<tr>
<td>2</td>
<td>5/15 (33%)</td>
<td>5</td>
<td>0/1 (0%)</td>
</tr>
<tr>
<td>3</td>
<td>2/5 (40%)</td>
<td>6-11</td>
<td>0/0 (0%)</td>
</tr>
</tbody>
</table>

Examples of patients with the AAD in the second dx position are shown in Table 3.

<table>
<thead>
<tr>
<th>AAD Dx position</th>
<th>Life-threatening AAD or history of AAD?</th>
<th>Primary Dx</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>786.5 chest pain</td>
<td>790.7 bacteremia</td>
</tr>
</tbody>
</table>

Conclusions
Our findings suggest that patients with an AAD dx in the second or third dx position may have life-threatening AAD but are more likely to have a previous history of AAD. The availability of large clinical datasets can facilitate research that is otherwise impractical, but accurate case definitions are required for appropriate use of these data.

References
Logical Rules for the Automated Ordering of Isolation Precautions

Jacob S. Tripp, PhD¹, Cherie Frame, RN, MSN, CIC², Crissy Elliott, MBA, MT (ASCP), CIC², Gisele G. Borsato, MS¹, Rebecca Farr, BS³
¹Homer Warner Center for Informatics Research; ²Infection Prevention; ³Nursing Operations;
Intermountain Healthcare, Salt Lake City, UT

Hospital-acquired infections (HAI) are one of the most common adverse events experienced by hospitalized patients. These infections include Methicillin-resistant Staphylococcus aureus (MRSA), Vancomycin-resistant enterococci (VRE), and other multidrug resistant organisms (MDRO), as well as many other less well-known diseases. The Centers for Disease Control and Prevention (CDC) maintains a list of isolation guidelines for nearly 250 infectious diseases.¹ As the numbers and types of infectious diseases are varied and often rare, Intermountain Healthcare has a history of using Computerized Decision Support (CDS) to aid clinicians in selecting appropriate isolation precautions for patients. In 2015, in connection with the implementation of a new electronic health record (EHR) installation, a team of infection control practitioners and engineers developed a set of logical rules to determine when patients needed to be placed on isolation precautions, and to automatically place orders for these isolation precautions, attempting to account for each disease in the CDC isolation guidelines.

These rules make use of a variety of clinical data elements to determine when isolation precautions are needed and the type of isolation to be ordered. The data elements used include a patient’s problem list, diagnoses, test orders, lab results, and data charted by nurses. Some examples of the rules are rules that: check a patient’s problem list upon admission for a history of MDRO infections; order isolation for patients who are diagnosed with hepatitis A or E and who also have evidence in their chart of having diarrhea or incontinence; order isolation precautions for pediatric patients for whom a rapid RSV (respiratory syncytial virus) test has been ordered; isolate incontinent patients with positive stool culture results; order isolation for patients with recent travel history to West Africa; isolate patients for whom nurses have charted evidence of rash, headache, fever, and photophobia, as this combination of symptoms may be an indication of meningitis. As of March 3, 2016, 113 logical rules for ordering isolation precautions had been created. Rules are continuously created and refined based on clinician feedback.

As a preliminary evaluation of these rules and the automated orders, a four-month sample of isolation orders at the largest hospital currently using them was studied. Between November 1, 2015 and February 29, 2016, there were 5030 orders for isolation precautions placed, with the most common reasons for isolation being MDRO-related (54% of orders). While the rules are intended to cover any situation when a patient would need to have isolation precautions ordered, nurses using the system can still order isolation manually any time they feel that it is needed. However, manual isolation orders made up only 3% of all isolation orders during the study period. While some isolation precautions are only needed for a limited time, and it is natural for them to be discontinued at some point during the patient’s hospital stay, 511 (10%) of the orders for isolation precautions were discontinued within 24 hours of their being ordered. Some of these orders, tied to lab test orders, were likely discontinued when those tests came back negative, but some of the orders may have been discontinued because they were unnecessary or inappropriate. Further analysis of discontinued orders may give helpful information in refining the rule logic.

This change, going from 100% of isolation orders being user-generated to 97% of isolation orders being system-generated, represents a major change in infection prevention workflow and merits further study to more fully understand its impact on HAI and other aspects of infection prevention.

References

Intravenous immunoglobulin stewardship using an electronic tool
Demetra Tsapepas, PharmD1,2, Hojjat Salmasian, MD MPH PhD1,2, David K. Vawdrey, PhD1,2
1NewYork-Presbyterian Hospital, New York, NY; 2Columbia University, New York, NY

Introduction
Healthcare organizations utilize electronic clinical decision support (CDS) to optimize medication use and safety. Preconfigured order sets may facilitate standardized treatment, encourage adherence to a particular care pathway, and optimize medication use--particularly for expensive medications. Intravenous Immunoglobulin (IVIG) has powerful therapeutic effects and is used across a wide spectrum of disease states. IVIG is a costly medication (several hundred dollars/dose) and its broad use constitutes a large expenditure for many healthcare organizations. Carefully implemented order sets may encourage evidence-based or institution-specific care through the influence on provider behavior.

Methods
In phases, an IVIG order set was designed and implemented in our electronic health record (EHR) during 2013. In the first phase, prescribers recorded an indication for IVIG in a free-text format. This crowd-sourced list of indications was then used in the second phase to develop a coded list of appropriate indications for IVIG. The order set would automatically calculate the appropriate dose of IVIG for 53 distinct conditions, adjusting for patient’s weight as needed. Although prescribers were allowed to place orders for IVIG through other routes (e.g., direct ordering for pediatric patients), the order set was the primary method for placing IVIG orders. Each condition was listed for prescribers to choose through a drop-down menu which subsequently populated the medication form including recommended dosage. Prescribers could manually override the recommended dosage while placing the order. We measured the utilization of the order set, and the rate of compliance with the guideline-recommended dosage, by querying 2.5 years of data from the EHR system. The study was conducted at NewYork-Presbyterian Hospital/Columbia University Medical Center and was approved by the institutional review board.

Results
During the 2.5-year study period, 1,039 orders were placed for IVIG, of which 814 were placed through the order set (78%). We limited our study sample to those where the patient’s height and weight was available at the time of order entry (N=724). The most common indications selected were various forms of transplant rejection, followed by Guillain Barre Syndrome, idiopathic thrombocytopenic purpura (ITP), and hypogammaglobulinemia. Using a tolerance range of ± 30% of the recommended dose, only 8% of orders placed for one group of indications (hypogammaglobulinemia) were for higher doses than recommended, although a notable fraction of orders were at a dosage lower than recommended (31%). Further analysis confirmed that overdosing was uncommon in orders placed for other indications using the order set.

Conclusion
An iteratively designed order set was commonly used by clinicians placing orders for IVIG and was associated with good compliance with the guideline recommended dosage. Overdosing of IVIG was rare in orders placed using the order set, although under-dosing remained relatively frequent. Our future research includes qualitative analysis of these orders to understand the reasons for noncompliance, an evaluation of the impact of underdosing on patient outcomes, and development of further enhancements to the order set accordingly.

References
Feature Engineering Alternatives for Natural Language Phenotype Classifiers
Clayton A. Turner, Paul E. Anderson, PhD, Alexander D. Jacobs, Cassios K. Marques, Jihad S. Obeid, MD
College of Charleston, Charleston, SC, Medical University of South Carolina, Charleston, SC

Introduction: Mining electronic health records (EHR) for known disease phenotypes allows automated identification of patients in the absence of International Classification of Diseases 9th revision (ICD-9) codes. This phenotype classification task can be automated using Natural Language Processing (NLP) techniques along with pattern recognition machine learning (ML) algorithms. The key to finding these patterns is translating plain text into quantifiable entities that could be used as features for ML. The classic technique for this is bag-of-words (BOWs). Another well documented method is to use UMLS Concept Unique Identifiers (CUIs) mapping with ML. The aim of this research is to evaluate the performance of these two approaches with various classifiers to identify patients with Systemic Lupus Erythematosus (SLE). A sub-aim is to determine which classification algorithms are better suited for each observed methodology, BOWs and CUIs.

Methods: With IRB approval, we obtained clinical notes for 322 patients with SLE diagnosis and 340 controls from the Rheumatology Clinic, a gold standard created by clinicians at MUSC. We filtered the notes for all patients, controls and SLE-diagnosed patients, to those only containing the partial word “rheumatol”, in order to enrich the text with more relevant notes to SLE (48,313 post-filtering). Classification using ICD-9 codes was used as a baseline metric. We created data manipulation scripts to transform the text notes to BOWs. The cTAKES/YTEX pipeline was used to produce CUIs. The variable importance extracted from the Random Forests classifier was utilized for feature selection. The output was subjected to ML algorithms using customized Theano and scikit-learn libraries: Neural Networks (NN) with 1 hidden layer, Random Forests, Naive Bayes and Support Vector Machines (SVM) with a radial basis function kernel. Performance was measured by calculating accuracy through cross validation and an external test set. The cross validated set contained its own training and test set (90%/10%) and the external set was created through a random sampling of 50 controls and 50 SLE patients.

Results: Table 1 shows the results of various combinations of approaches and classifiers along with the results based on ICD-9 codes.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Data Type</th>
<th>ICD-9</th>
<th>NN</th>
<th>Random Forests</th>
<th>Naive Bayes</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ext. Acc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>N/A</td>
<td>BOW CUIs</td>
<td>BOW CUIs</td>
<td>BOW CUIs</td>
<td>BOW CUIs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89.66</td>
<td>94.14</td>
<td>95.17</td>
<td>95.35</td>
<td>90.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>92.00</td>
<td>82.00</td>
<td>97.00</td>
<td>92.00</td>
<td>84.00</td>
</tr>
</tbody>
</table>

Discussion: Currently cohort identification in EHRs is most commonly achieved using ICD-9 billing codes. In our dataset we used ICD-9 as a baseline and were able to achieve 90% accuracy. The NN performance was much better using CUIs, according to the cross validation and external test set accuracies. This is presumed to be due to CUIs providing more refined and contextual patterns than simple word frequencies in BOWs for the NN to recognize. We hypothesize that with more training data and a deeper network, the NN would perform more effectively with BOWs. The relatively good performance of Random Forests with BOW is consistent with historical performance of these classifiers with BOW. It is important to note that in our experience generating the CUIs was much more resource intensive than generating BOWs, rendering this Random Forest performance with BOW fairly desirable. The “rheumatol” filtering on the notes removed a large amount of noise from the data, allowing each classifier to train more effectively. The naive bayes algorithm was included as a baseline as it operates on the assumption that features are independent when, in this context, words and concepts together are very dependent upon each other. Lastly, SVMs are often an effective classifier for convoluted patterns in textual datasets.

Conclusion: Our results suggest that a shallow NN with CUIs performed best as a classifier, however the overhead in generating CUIs has to be balanced with the almost equally good performance using the Random Forests classifier with BOWs, though the Random Forests did not perform as well on the external set. Our examinations on the feature sets generated by these techniques suggest CUIs are more generalizable than BOWs, as well. Future considerations include: examining the performance of these classifiers on different portions of clinical notes e.g. chief complaint, history of present illness, family history, assessment and plan; and applying these algorithms to phenotyping in other disease states for a variety of uses cases, such as clinical trials eligibility and genotype-phenotype studies.

Acknowledgements: This work is funded by the National Institutes of Health (NIH) Grant #s P60AR062755 and UL1TR000062, the Medical University of South Carolina, the College of Charleston and the SmartState Program in SC.
Implementation Strategies to Address Information-related Barriers for Community Pharmacies Participating in a Statewide Pharmacy Care Management Network

Kea Turner, MPH, MA1, Christopher M. Shea, PhD, MA, MPA1
1. University of North Carolina-Chapel Hill, Gillings School of Global Public Health, Chapel Hill, NC

Introduction. Approximately 250 community pharmacies in North Carolina are participating in a pharmacy care management network designed to improve care quality and reduce cost of care.1 Pharmacies in the network are implementing changes needed to provide enhanced clinical services, such as comprehensive assessments of medications, and coordinate care with primary care physicians and other providers. Previous work has identified information-related barriers pharmacies in the network are facing, such as challenges with learning a new documentation system and integrating documentation within workflow.2

Objective. The goal of this study is to systematically document implementation strategies that network staff and program collaborators are employing to assist community pharmacies with addressing information-related barriers to implementing enhanced clinical services and improving care coordination. This documentation will facilitate evaluation of the implementation strategies and dissemination of effective strategies to other networks.

Methods. We are in the process of conducting 15 one-hour interviews with network staff and program collaborators (i.e., network operational staff, university research team members) who are employing implementation strategies to assist participating community pharmacies with addressing information-related barriers. Interviews are being conducted using a semi-structured interview guide based on Powell et al.’s compilation of implementation strategies3, which defines 73 possible implementation strategies, and Proctor et al.’s guidance for specifying implementation strategies, which provides guidance for documenting the strategies in detail using a template.4

Results. Preliminary results indicate that community pharmacists experience information-related barriers, such as difficulty locating patient data and obtaining medication information from primary care providers. To address these barriers, network staff and program collaborators have employed a number of implementation strategies, such as developing and implementing tools for quality monitoring, developing and organizing quality-monitoring systems, conducting educational outreach visits, assessing change readiness, and conducting audit and feedback.

Discussion. The interviews have yielded detailed information about the implementation strategies being employed. Also, preliminary feedback from interviewees suggests that strategies such as audit and feedback, tools for quality monitoring, and educational outreach visits may facilitate effective information use by community pharmacies. However, systematic evaluation of each strategy is needed to determine impact on implementation of the enhanced services. The documentation conducted during this study will facilitate evaluation of the strategies.

Conclusion. When introducing clinical services within community pharmacies, documenting implementation strategies employed to address information-related barriers is important for evaluating the implementation approach, promoting efficient use of implementation resources, and ultimately supporting successful practice change.

About the authors. Dr. Christopher Shea and PhD student Kea Turner are currently working with Community Care of North Carolina to evaluate implementation of a pharmacy care management network in NC.

References
Understanding Technology Requirements to Support Chronic Disease Care: The Longitudinal Care Plan Cycle

Kim M. Unertl, PhD, MS1, Christopher L. Simpson, MA1, Laurie L. Novak, PhD1
1Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Introduction

Optimizing chronic disease care is a substantial and growing challenge for the healthcare system, with a need to shift from episodic acute care delivery approaches towards longitudinal perspectives on health. Transformational approaches such as the Chronic Care Model include health information technology as important components in chronic disease management. However, too little is known about specific technology requirements to support chronic disease work processes and information needs, a major barrier to designing usable and useful technology.

Methodology

The study was set in an academic-community collaboration that had a goal of creating a Medical Home Model for individuals with sickle cell disease (SCD) from birth through adulthood. Researchers collected data from May 2012-July 2014 using non-participant and participant observation, informal interviews, and artifact and spatial data collection. Data collection was open-ended and continued until a point of data saturation was reached. Data were analyzed using a Grounded Theory approach, with iterative cycles of progressively more focused coding.

Results

As data collection progressed, strong themes emerged around longitudinal care planning. Managing the health of an individual with a complex chronic disease is a shared process among the individual, informal caregivers, and healthcare teams. We observed the development of collaborative plans for health management, key work activities that occurred without formal comprehensive support from the existing health information technology infrastructure. Based on our data we developed a six-stage collaborative work process model, the Longitudinal Care Plan Cycle (Figure).

Discussion and Conclusion

Developing and acting on a longitudinal care plan for individuals with a chronic disease is more than a single event. Instead, the process is complex and continuous and involves many actors. The Longitudinal Care Plan Cycle supplements existing approaches such as the Chronic Care Model and the Medical Home Model to assist with identifying specific process-oriented technology requirements. Requirements for health IT to support longitudinal care plan processes include: communication tools for boundary-crossing collaboration, temporally-oriented longitudinal planning elements, data visualization of health status over time, and collection and analysis of patient-specific barriers to health management.

Figure. The Longitudinal Care Plan Cycle
Structured Information Displays for the Comparison of Clinical Trials

Prasad Unni¹, MD, Jiantao Bian¹, MS, Charlene Weir¹,², BS, PhD, RN, Guilherme Del Fiol¹, MD, PhD

¹Biomedical Informatics, University of Utah, Salt Lake City UT; ²George E. Whalen Veterans Affairs Medical Center, Salt Lake City, UT

Introduction

Clinicians raise information needs during patient consultations that are often not met. Evidence from randomized controlled trials (RCT) is an important source to help clinicians meet their information needs. Yet, this evidence source is underutilized due to the significant effort required to judge relevance and interpret results of RCTs for a specific patient. Murlow et al. proposed that tabular displays can help systematic reviewers, patients, and providers integrate research findings into practice¹.

In our previous study that investigated alternative information displays for RCT reports, we found that clinicians favored a structured, tabular display vs. narrative abstracts². However, clinicians still found the tabular display somewhat overwhelming. To address this issue, we designed different displays focused on two steps in the RCT information seeking process: relevance judgment and interpretation of study results.

Methods

We followed a user-centred design process, which included multiple iterations within our team of students, informatics researchers and clinicians. We used NinjaMock®, a low-fidelity prototyping tool to design, and HTML, CSS, JavaScript to build a fully interactive prototype. After several cycles of feedback and redesign, we exposed the prototype to a group of clinicians for broader informal feedback. For the prototype, we manually extracted relevant information from a set of clinical trials and “hard-coded” into a JSON file used by the prototype.

Results

The resulting prototype (Demo link: https://goo.gl/Ins9ig) provides users with four layers of interaction. The first three layers are designed to facilitate efficient relevance judgment and the fourth layer provides a tabular comparison for the interpretation of study results. The first layer clusters search results into studies that have similar patient characteristics and interventions. The second and third layers are used to “zoom into” a selected cluster and select individual RCTs for comparison. The final layer (Figure 1) uses a tabular format to present RCT details, which can be compared among RCTs. Similar attributes across RCTs, are displayed in the same row to allow direct comparison.

Discussion

The proposed tool is a departure from traditional approaches to the visualization of search results for RCT reports. Unlike narrative abstracts, our information display was explicitly designed to facilitate two distinct and critical steps in the information-seeking process: relevance judgement and study interpretation. After further refinement based on clinician input, we plan to conduct a formative evaluation of the information display, comparing with traditional narrative formats.

References


Acknowledgement: Supported by grant 1R01LM011416-01 from the National Library of Medicine.

Figure 1: Comparison table with selected studies arranged in columns.
Temporal Annotation of Swedish Intensive Care Notes

Sumithra Velupillai, PhD¹, Rebecka Weegar, MSc², Maria Kvist, PhD²,³
¹KTH, Sweden; ²Stockholm University, Sweden; ³Karolinska University Hospital, Sweden

Abstract
We describe the creation of a corpus of Swedish intensive care unit (ICU) notes annotated for temporal expressions.

Introduction
To better understand disease and care progression, automated extraction of time expressions can help build models for temporal inference. TimeML¹ is a model for semantic annotation of text for enabling temporal reasoning, which has been adapted to the clinical domain for English²,³. Our long-term goal is to develop a state-of-the-art temporal reasoning system for Swedish clinical text that extracts temporal relations between events and time expressions.

Methods and Materials
Clinical notes from an ICU in Stockholm, Sweden were used. The HeidelTime⁴ system was adapted to develop Swedish clinical time expression (TIMEX3) resources. An iterative process was employed: 1) automated pre-annotation, 2) manual identification of errors and missing expressions (domain expert, A1). Inter-Annotator Agreement (IAA) was calculated on a final test set (17 notes), annotated by A1 and A2 (NLP researcher), using precision, recall and F-measure, exact and overlap span, per TIMEX3 type and overall micro- and macro-average.

Results
Over 1800 TIMEX3s were identified (1834 A1, 1864 A2). One note=one patient’s entire ICU documentation. Overall micro-average IAA is high (86% F1), Table 1. Date is the most common type. Prepostexp (expressions about operations) yields highest IAA, indicating unambiguous expressions. Frequency results in most discrepancies. The notes contain a wide variety of time information, i.e. the resulting terms are likely useful also in non-ICU notes.

Table 1. Results: TIMEX3 annotations in Swedish ICU notes compared between annotators A1 and A2.

<table>
<thead>
<tr>
<th>TIMEX3</th>
<th>exact</th>
<th></th>
<th></th>
<th>overlap</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-measure</td>
<td>Precision</td>
<td>Recall</td>
<td>F-measure</td>
</tr>
<tr>
<td>Date</td>
<td>0.83</td>
<td>0.90</td>
<td>0.86</td>
<td>0.89</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>Duration</td>
<td>0.63</td>
<td>0.57</td>
<td>0.60</td>
<td>0.84</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.73</td>
<td>0.34</td>
<td>0.47</td>
<td>0.81</td>
<td>0.39</td>
<td>0.53</td>
</tr>
<tr>
<td>Prepostexp</td>
<td>1.00</td>
<td>0.94</td>
<td>0.97</td>
<td>1.00</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>Time</td>
<td>0.90</td>
<td>0.95</td>
<td>0.92</td>
<td>0.93</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Overall (macro/micro)</td>
<td>0.82/0.85</td>
<td>0.74/0.87</td>
<td>0.76/0.86</td>
<td>0.89/0.91</td>
<td>0.82/0.92</td>
<td>0.84/0.91</td>
</tr>
</tbody>
</table>

Discussion and Conclusion
We present a Swedish clinical corpus annotated for TIMEX3s. Our IAA results are comparable to previous studies, (73%/87% (exact/overlap)² and 80%/³ F1). We have created Swedish lexical resources with clinically specific time expressions that will be useful for the development of clinical temporal reasoning systems also in other domains.

Acknowledgements
This work was partially funded by the Swedish Research Council (350-2012-6658) and Nordic Information for Action eScience Center (NIASeC); a Nordic Center of Excellence, Nordforsk (62721).

References

¹ Approval by the Regional Ethical Review Board in Stockholm: 2012/834-31/5

1622
Cajun Codefest 4.0 on SMART-on-FHIR apps for Diabetes

Kavishwar B. Wagholikar, MBBS PhD 1,2, Rahul Jain, MPH CPHIM3, Eliel Oliveira, MBA MS3, Henry Chu, PhD PE4, Harshal Shah 5, Joshua Mandel, MD 1,6, Jeffery Klann, PhD 1,2, Sohail Rao, MD, MA, D.Phil 7, Kenneth D. Mandl, MD MPH 1,6, Shawn N. Murphy, MD PhD 1,2, Thomas Carton, PhD, MS 3

1Massachusetts General Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Louisiana Public Health Institute, New Orleans, LA; 4University of Louisiana at Lafayette; 5Persistent Systems, India; 6Children’s Hospital Boston, Boston, MA; 7Ochsner Health System, LA

Introduction

Clinical Data Research Networks (CDRNs) sponsored by Patient-Centered Outcomes Research Institute (PCORI) widely use the infrastructure developed by the ‘informatics for integrating biology and the bedside’ (i2b2) project to store data. i2b2 has developed a FHIR interface, to facilitate development of reusable apps that can be shared across the CDRNs. However, there is a lack of expertise in the clinical IT community to develop apps using SMART on FHIR.

In order to spur the development of programming skills for these upcoming technologies the authors collaborated to conduct Cajun Code Fest 4.0 (CCF), posing a challenge for development of Diabetes apps. We constructed a platform for CCF in the REACHnet data center hosting data for 300 real patients diagnosed with Diabetes. This data was anonymized by date-shifting and by substituting the patient identifier. A FHIR endpoint was made available to CCF participants on their signing of a confidentiality agreement with REACHnet. Education videos were developed which provide an introduction to the FHIR api. These resources are published online and are freely available at http://www.cajuncodefest.org/index.php/event-info/data-release. For executing the Code Fest, we created parallel tracks for engagement and for infrastructure.

Engagement

REACHnet partnered with CCF on the fourth iteration of their annual health coding competition. We gave seminars on health informatics at UL Lafayette’s Informatics Research Institute to prospective CCF participants in October and November. Next, we organized a connect-a-thon at Ochsner Health System in New Orleans, LA on December 11th, 2015 to empower CCF and other regional developers. To kick start the CodeFest, subject matter experts engaged 40 participants on a range of topic including the learning health system, SMART Health IT technology, FHIR, HAPI, patient centered outcomes research, and how to access REACHnet’s data infrastructure to compete in Cajun Code Fest. The event was recorded and shared on social media. Connect-a-thon attendees were encouraged to compete in CCF by submitting a concept paper in February of 2016 with final judging being in April of 2016. Attendees were provided confidentiality agreement to access the FHIR API at REACHnet's data center. A REACHnet Clinical Innovation Committee was created to facilitate app ideation and link CCF competitors with subject matter experts in clinician science, medical care, and population health; and to support award adjudication.

Infrastructure

REACHnet deployed an API for the development of reusable apps by CCF that draws from the PCORnet Common Data Model V 2.0. REACHnet availed its data center and created an i2b2 instance with 300 real patients. The dates were shifted at the patient level by REACHnet to deidentify the data in consultation with national privacy experts. The FHIR cell i2b2 plugin was installed in REACHnet’s i2b2 instance to support SMART on FHIR. A demo application was created to validate the functionality of a reusable app that leverages SMART on FHIR on i2b2 in accessing the PCORnet CDM. OAuth user access tokens were created to authenticate calls to the FHIR API.

It is expected that 5-7 teams will move on to be judged for a CCF award leading to entrepreneurial and incubation services. The competition will yield apps that are of value to health systems, researchers, and patients. The educational artifacts created for the Code fest participants will serve as a useful starting point for developers.

Acknowledgements: Research, dissemination, and engagement for the REACHnet connect-a-thon and its associated data infrastructure was supported in part by grant CDRN-1306-04864 from the Patient Centered Outcomes Research Institute, and partly by R00LM011575 from NLM, and R01GM104303 from NIGMS.
Planning a Learning Health System Evaluation: The Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATHS) Initiative

J.S. Wald, MD1, J.E. Richardson, PhD1, M. Ashok, PhD1, J.R. Webb, MS1, B.H. Blumenfeld, MD1, A. Ortiz, MS1, R. Bailey1, G. Phillips, PhD2, J.R. Williams, PhD2, R. Rudick, MD2

1RTI International, Research Triangle Park, Raleigh, NC; 2Biogen, Cambridge, MA

Introduction
Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATHS) is a Learning Health System demonstration project sponsored by Biogen in collaboration with 10 healthcare institutions in the United States and Europe. MS PATHS aims to integrate technology into routine MS care to standardize the collection of clinical data in order to learn from routine clinical practice and generate insights that one day may be used to inform decision making at the point of care. MS PATHS leverages a new iPad-based patient self-administered assessment tool, the Multiple Sclerosis Performance Test (MSPT), to objectively quantify major motor, visual, and cognitive symptoms; implements standardized acquisition sequences of magnetic resonance imaging to enable quantitative image analysis; and employs other technologies that form an informatics infrastructure for de-identifying and pooling patient data across multiple institutions. RTI International is conducting a program evaluation of the MS PATHS demonstration project funded by Biogen. This poster presents the comprehensive plan for the program evaluation of this important and complex initiative.

Methods
We developed an evaluation plan (EP) using an adapted version of the Consolidated Framework for Implementation Research1. EP development included: 1) a literature review to identify relevant outcomes and measures, 2) discussions with MS PATHS sponsors at Biogen and leadership at participating institutions, and 3) a measurement strategy using an evaluation planning matrix (EPM) that maps program goals and success indicators to specific metrics. Using the EPM, we developed surveys, interview guides, and tracking metrics, leveraging established instruments where possible. The draft program EP was reviewed by six key opinion leaders unaffiliated with the project, Biogen sponsors, and site leadership, who were invited to review and critique the draft. The RTI institutional review board confirmed this work was considered program evaluation, not research.

Results
The EP is driven by three overarching evaluation questions: 1) How successful is the MS PATHS implementation at each institution? 2) What collaboration is taking place among network institutions to support research, improve implementations, and promote mutual learning? 3) How is research productivity growing as a result of MS PATHS, and what would accelerate further growth? The EP calls for mixed methods data collection using surveys, semi-structured interviews, site visit observations, document analyses, and system tracking data. Data collection at each institution is timed to coincide with early implementation (2-3 months post-go-live) and routine use (after 8-10 months), with annual follow-up visits in the remaining 36 month period.

Constructs guide data collection and analysis. Institution implementation constructs include alignment of MS PATHS with institution clinical and business priorities; access to and perceived value of training; the reach of the tools (# of patients, # of assessments); value of MSPT data in direct clinical care; and the patient experience of care. Network constructs include collaboration across institutions to accelerate problem-solving; data coordination; and effective governance. Research productivity constructs include growth in research capacity (e.g. data contributions); indicators of high data quality (e.g. accuracy, completeness, availability); research activity (e.g. patient cohort queries, extracts, protocols); and dissemination (e.g. presentations, publications).

Discussion
The EP highlighted several timing-related items to consider during data collection and analysis. First, data collection at each institution should be timed to align site implementation phase if possible. Second, research activities such as data capture precede other indications of research productivity, so early assessment must take this into account. Third, the EP itself is likely to evolve as additional institutions, including European institutions, are live. Fourth, site implementations are a moving target, suggesting that multiple data collection time points will be helpful.

References
Healthcare Providers’ Usage of an Interactive Case Simulation Tool for HIV Patient Mental Health Screening in a Statewide Clinical Education Program

Dongwen Wang, PhD
Arizona State University, Scottsdale, AZ

Abstract

We conducted a pilot study to analyze the usage of an interactive case simulation tool (ICST) by healthcare providers for mental health screening in HIV patients. We identified specific usage patterns and verified their generalizability. Future development of ICSTs should consider such patterns for customized tool design and effective dissemination.

Introduction

Dissemination of clinical evidence to healthcare providers is an essential requirement for translation of biomedical research into patient care. We have demonstrated in previous studies that: (a) interactive case simulation tools (ICSTs) can be used as an effective channel for knowledge dissemination; and (b) there are specific patterns in clinicians’ usage of ICSTs, as shown in a case study of an ICST for insomnia screening and treatment in HIV patients. Here we report a study: (a) to analyze clinicians’ usage patterns on a different ICST for mental health screening in HIV patients; and (b) to verify the common ICST usage patterns by comparing the findings from this study with the previous one on insomnia ICST. The results from this study will help identify the generalizable usage patterns across different ICSTs, and direct future development of such tools for more effective and targeted knowledge dissemination.

Methods

We conducted the study within the New York State (NYS) HIV-HCV-STD Clinical Education Initiative (CEI) online education program (www.ceitraining.org). Both ICSTs in the current and the previous studies were part of CEI’s repository of online educational resources. To ensure a fair comparison, we selected the same study period (from April 3, 2012 to October 15, 2012) and queried the usage tracking database for data collection. Parameters for usage pattern analysis included: (a) new vs. returning users, (b) access through web browsers vs. native apps, and (c) access from large-screen equipment (desktops, tablets, etc.) vs. small-screen hand-held devices (smartphones, iPods, etc.). Usage analyses of specific ICST sections were based on the same classifications in the previous study, i.e., “recommendations”, “sample cases”, and “user-defined cases”. We used Chi-Square test to examine the statistical significance between the usage groups.

Results

We recorded a total of 386 user visits to the mental health screening ICST during the study period. Among them, 381 (98.70%) were from new users and 5 (1.30%) from returning users. Regarding the platform, 283 (73.32%) visits came from native apps, while the remaining 103 (26.68%) were through web browsers. In terms of screen size, 144 (37.31%) visits originated from large-screen equipment and 242 (62.69%) were from small-screen devices. Pattern analyses found that: (a) visits to “sample cases” and “user-defined cases” were more likely from small-screen users (p=0.001), (b) visits to “sample cases” and “user-defined cases” were more likely from app users (p<0.001), and (c) small-screen users were more likely to access the resources through native apps (p<0.001). These findings reconfirmed the usage patterns identified from the previous study of the insomnia ICST. Although we also found that returning users were more likely to use large-screen equipment and to access the resources through web browsers, the sample size in specific usage context was too small to make any valid statistical inferences.

Conclusion

Initial analyses have shown that certain patterns of healthcare providers’ usage of ICST can be generalized to multiple clinical problems. Future development of ICSTs should consider these usage contexts and patterns for customized tool design and more effective knowledge dissemination.

Acknowledgement: This work is supported by grant #R24HS022057 from the Agency for Healthcare Research and Quality (AHRQ), and by contracts #C023557, #C024882, and #C029086 from NYS Department of Health AIDS Institute. The content is solely the responsibility of the author and does not necessarily represent the official views of the sponsors. We would like to thank Dr. Xuan Hung Le, Dr. Amneris Luque, and other CEI program staff for their contributions to this study. We would like to thank AHRQ and NYS program officers Marian James, Beatrice Aladin, Cheryl Smith, Howard Lavigne, Lyn Stevens, and Bruce Agins for their support.
Medication Recommendation for Chronic Diseases with Comorbidities Using Electronic Medical Records

Liqin Wang, MS¹,², Peter J. Haug, MD¹,²
¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA;
²Intermountain Healthcare, Salt Lake City, UT, USA

Objective: The goal of this study is to identify effective medications among a number of related drugs for chronic diseases with comorbidities, and to identify medications potentially causing adverse events when prescribing them to patient with comorbidities.

Background: Healthcare providers are likely to consider a number of drugs when developing a treatment plan for their patients. For example, diuretics, β-blockers, calcium-channel blockers, and ACE inhibitors, are classes of drugs used for treating heart failure. An important assumption is that medications may affect different subpopulations (e.g., patients with comorbidities) differently. FDA-approved indications or clinical practice guidelines may lack the details to determine how well the medications treat patients with a typical circumstances, especially special population like the elderly or those having multiple comorbidities, while practical data stored in the electronic medical records could be very useful by supplementing evidence missed from these sources. For chronic diseases, the treatment is usually long-term; a medication which is not effective in a certain subgroup will be stopped quickly and more effective medication will end up with more frequent prescriptions. This could statistically differentiate the effective medications from those not effective medications with well-defined diagnostic groups. It is important to identify the most effective medications for patient subgroups among those medications with similar indications in order to reduce poor medication adherence and potential adverse events. In this study, we intend to extract and sort medications for the treatment of chronic diseases with and without common comorbidities. We will identify those comorbidities that result significant differences in the patterns of prescription choice compared to patients with no such comorbidities. Identification of these comorbidity-associated differences may be useful in identifying the need for changes when prescribing medications to patients with those phenotypes.

Methods: In the initial stage of the study, we explored the chronic heart failure (CHF) population with different comorbidities using the data from the Intermountain Healthcare enterprise data warehouse (EDW). We first extracted a CHF-pertinent, pharmaceutical treatment vocabulary as a reference from a heart failure ontology.¹ We then identified a range of CHF-pertinent comorbidities from the biomedical literature and verified their frequency of occurrence in the CHF population from the EDW. For each comorbidity, we exploited association mining algorithms (e.g., support, chi-square, and interest) to extract and rank the medications related to the CHF with and without comorbidity. Extracted medications not directly included in or being members of the reference vocabulary were excluded from the analysis. Thereafter, we compared the rankings of CHF-pertinent medications for populations with and without certain comorbidity, and analyzed changes in the ranks of the same medications from the two subpopulations.

Results/Conclusion: We totally retrieved 186 pharmaceutical treatment terms as a reference. We initially tested our approach on five non-cardiac comorbidities: diabetes mellitus, chronic kidney disease, asthma, arthritis, and depression. We obtained two lists of ranked medications for each comorbidity: one is for CHF with comorbidity, and the other is for CHF with no comorbidity. Although the majority of the medications were ranked differently in two compared lists, furosemide was ranked as top 1 medication in most lists. By comparing the ranked medications between the paired lists, we identified some medications dropped significantly in their ranking from the subpopulation with comorbidities as compared to the non-comorbid subpopulation. There were (1) lisinopril, potassium chloride, spironolactone, and warfarin for the comorbidity of kidney disease; (2) isosorbide and clopidogrel for the asthma comorbidity; (3) metalazone, warfarin, isosorbide, and torsemide for the arthritis comorbidity; and (4) amiodarone, warfarin, isosorbide, and lisinopril for the depression comorbidity. These medications will be further analyzed to investigate their potential for adverse reactions in the affected subgroups. The lists of ranked medications from this study may prove to be useful for recommending pharmaceutical choices for CHF patients with comorbidities.

References

Discovering Associations between Problem List and Practice Setting

Liwei Wang, MD, PhD, Dingcheng Li, PhD, Majid Rastegar-Mojarad, MSc, Hongfang Liu, PhD

Department of health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, U.S.

Abstract

The problem list has a potential to help practitioners to provide customized care to a patient. However, it remains an open question on how to leverage problem lists in different practice settings for providers to provide tailored care. In this study, we aim to investigate the associations between the problem list and practice settings using the longitude data from Mayo Clinic EHRs, through natural language processing (NLP) and topic modeling techniques.

Introduction

The Health Information Technology for Economic and Clinical Health Act (HITECH) has greatly accelerated the adoption of electronic health records (EHRs) with the promise of better clinical decisions and patients’ outcomes. One of the core criteria for “Meaningful Use” of EHRs is to have a problem list which lists the most important health problems facing a patient. The problem list was first introduced by Weed in 1968 in his promotion for a Problem-Oriented Medical Record (POMR) 1. As important communication vehicles, problem lists are supposed to play an indispensable role in clinical practices. However, it remains an open question on how to leverage problem lists in different practice settings for providers to provide tailored care. In this study, we aim to investigate the associations between the problem list and practice settings using the longitude data from Mayo Clinic EHRs, through natural language processing (NLP) and topic modeling techniques.

Method

The collection of clinical documents used in our analysis consists of clinical notes for a cohort of patients receiving their primary care at Mayo Clinic, spanning a period of 15 years (1998–2013), and covering both inpatient and outpatient settings. Problems in those documents are generally itemized entries as either phrases (e.g., Allergic rhinitis/vasomotor rhinitis) or short sentences (e.g., Her asthma appeared to be very mild). We used the NLP method from our previous study to process the problem list, extract problems and normalize them to codes from the unified Medical Language Systems (UMLS). Then we kept the codes that mapped to the CORE Problem List Subset codes. According to the features of settings in Mayo Clinic notes and extended values of HL7/LOINC DO, we further aggregated them into 73 settings.

We used R package “topicmodels” to build our topic model. To choose the optimal topics in our dataset, we calculate the log likelihood values with the number of topics varied from 5 to 150 by 5, and then investigated the performance by comparing the log likelihood value, of which the highest indicates the optimal number of topics. And the result returns 65 as the optimal topics. Then we fit an LDA model with 65 topics using Gibbs sampling with a burn-in of 1000 iterations.

Initial Results

Figure 1 shows the histograms of the topics for all 73 settings. Fig. 1A shows the histogram of the frequency of the most likely topics for documents from 73 settings. The dataset is divided into 65 categories (topics), and the frequency (document number) difference of various topics could be observed. For example, for all documents, topic 15, 33 and 5 are most frequent. Topic 15 represents a theme of “patient laboratory tests”, topic 33 “surgery”, and topic 5 “pain”. Fig 1B shows the histogram of frequency of top 3 topics in 73 settings, and there are no significant frequency difference for most topics except topic 1 and topic 2.

Future work

We would choose representative problem concepts of each setting with statistical significance, and evaluate the accuracy in new clinical notes.

References

Auditing Tree-like Organ Systems in the FMA Using Network Motifs

Lucy L. Wang, MS\(^1\), Eli Grunblatt\(^2\), Mark E. Whipple, MD, MS\(^1,3\)

\(^1\)Department of Biomedical Informatics and Medical Education
\(^2\)Medical Scientist Training Program
\(^3\)Department of Otolaryngology-Head and Neck Surgery
University of Washington, Seattle, WA

The Foundational Model of Anatomy (FMA) ontology forms the basis of many biological and physiological models involving human anatomy.\(^1\) A graphical analysis of tree-like organ systems in the FMA reveals highly-recurrent network motifs. One such motif is the wing formation, consisting of a superclass anatomical entity and its relationship to its two subclasses in the left- and right-side body. This is shown in Figure 1a, where a series of arterial trunks have symmetric left and right subclasses. A double wing formation is a fully connected pair of wings, showing continuity between the two superclass nodes, and between the respective left and right nodes. This motif exists in FMA organ systems that display bilateral symmetry and connectivity (i.e. continuity, flow, or attachment).

Figure 1. (a) A portion of the arterial tree from the FMA, (b) a double wing formation and five broken double wing formations (i-v). Black lines show connectivity relationships; gray lines show subclass relationships. Red nodes are superclasses; blue nodes are left and right subclasses.

Figure 1b documents several broken double wing formations, which can be identified computationally by searching the connected graph of each organ system. From the FMA, we generated organ system views of three vascular trees: the systemic arterial, systemic venous, and lymphatic systems. Unique instances of broken formations are detected for each vascular tree (Table 1) and are especially prevalent in the lymphatics.

Table 1. Broken double wing formations for each of the three FMA vascular tree organs.

<table>
<thead>
<tr>
<th>FMA organ tree</th>
<th>i</th>
<th>ii</th>
<th>iii</th>
<th>iv</th>
<th>v</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>arterial</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>venous</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>lymphatic</td>
<td>61</td>
<td>74</td>
<td>12</td>
<td>5</td>
<td>8</td>
<td>160</td>
</tr>
</tbody>
</table>

This work builds on the use of bilateral symmetry\(^2\) and organ-specific applications\(^3\) to audit the FMA, and focuses on cases of recurring graph motifs. Inconsistencies in content can be discovered by applying this technique to relevant organ systems, allowing for increasingly computational audits of the FMA ontology. This method may also be generalizable to other resources, such as the anatomical portion of SNOMED-CT and the cross-species anatomical ontology Uberon, or another ontology in which structural relationships are reflected in the structure of the resource.

Research reported in this abstract was supported by the National Library Of Medicine Training Grant T15LM007442.

References

A Part-Of-Speech Weighting Scheme for Medical Information Retrieval

Yanshan Wang, PhD1, Stephen T. Wu, PhD2, Dingcheng Li, PhD1, Hongfang Liu, PhD1
1Mayo Clinic, Rochester, MN; 2Oregon Health & Science University, Portland, OR

Introduction

Over the past decade, the amount of health information has increased exponentially in the form of computerized Electronic Health Records (EHRs). The second use of EHRs has facilitated healthcare solutions and clinical decision making, from which physicians and patients have been tremendously benefited in the clinical practice. As the EHR repository becomes larger, however, it becomes more intractable to search the preferred health information in the current systems. Information retrieval (IR), an interdisciplinary research field of information science and computer science, has become a crucial technology to find textual information. Since the EHRs are also narrative textual data, IR has become an essential part of the modern EHR systems.

Methods

In this study, we propose an automatic Part-of-Speech (POS) based query term weighting scheme for two widely used IR models, Bag-of-Word (BoW) model and the Markov Random Field (MRF) model, which are named POS-BoW and POS-MRF, respectively. The motivation is that the Natural Language Processing (NLP) techniques, specifically, POS tagging, would strengthen the conventional IR models to improve the retrieval performance in the clinical domain. The posterior ranking functions of POS-BoW and POS-MRF are defined as:

\[ r_{\text{POS-BoW}}(Q, D) = \sum_{\tau \in \mathcal{T}} \sum_{q \in \tau} \lambda_q f(c, q, D) \]

and

\[ r_{\text{POS-MRF}}(Q, D) = \sum_{c \in (T, O, U)} \theta_c \sum_{\tau \in \mathcal{T}} \sum_{q \in \tau} \lambda_q f(c, q, D) \]

respectively, where \( D \) denotes document set, \( Q \) query set, \( \mathcal{T} = \{1, 2, ..., \tau\} \) the POS category set, \( c \in \{T, O, U\} \) is a set of cliques, \( T \) the single term cliques, \( O \) the ordered term cliques, \( U \) the unordered term cliques, \( \lambda = \{\lambda_1, \lambda_2, ..., \lambda_\tau\} \) is the weight vector for terms with distinct POS categories, and \( f(c, q, D) \) the feature functions over clique \( c \) for query \( q \). Figure 1 depicts an example of graphical models of POS-BoW and POS-MRF with three query terms. In order to decide the weight, we employ a supervised machine learning algorithm. This algorithm is a cyclic coordinate method where golden section line search algorithm is applied along each coordinate to optimize the Mean Average Precision (MAP) defined objective function.

Results

The effectiveness of the proposed approach was verified on a twofold cross validation using TREC 2011 and 2012 Medical Records tracks. Language Models with Dirichlet Smoothing was used in the feature function. The experimental results are listed in Table 1.

Table 1. Average results on the TREC 2011 and 2012 Medical Records tracks.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP (impr)</th>
<th>bpref (impr)</th>
<th>P@10 (impr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.257 (-)</td>
<td>0.355 (-)</td>
<td>0.424 (-)</td>
</tr>
<tr>
<td>POS-BoW</td>
<td>0.279 (8.6%)</td>
<td>0.363 (2.3%)</td>
<td>0.430 (1.4%)</td>
</tr>
<tr>
<td>MRF</td>
<td>0.265 (-)</td>
<td>0.377 (-)</td>
<td>0.468 (-)</td>
</tr>
<tr>
<td>POS-MRF</td>
<td>0.318 (20%)</td>
<td>0.404 (7.2%)</td>
<td>0.504 (7.7%)</td>
</tr>
</tbody>
</table>

Conclusion

The experimental results are consistent with our hypothesis that the NLP techniques, specifically POS tagging, could enhance IR models and improve the retrieval performance in the clinical domain.
Personal Health Record Self-Scheduling by Patients Decreases No-Show Rates and Operational Efficiency in Primary Care and Specialty Follow-Up Visits

Michael A Wassef, MD,1 Bret Wagner, BS,3 Joshua Russow, BS,3 David C Kaelber, MD, PhD, MPH1,2

1School of Medicine, 2Center for Clinical Informatics Research and Education, The MetroHealth System, Case Western Reserve University, Cleveland, OH, 3Huron Healthcare, Huron Consulting Group, Madison, WI

Background
The MetroHealth System (MHS) (Cleveland OH) has over 40% of its socio-economically diverse patients signed up for its personal health record (PHR) tethered to its Epic (Verona WI) electronic health record (EHR). One of the features in the PHR is patient appointment self-scheduling. Self-scheduling allows the patient to select and reserve the date, time, and location for an appointment that is most convenient for the patient in real-time. MHS deployed PHR self-scheduling for primary care follow-up appointments in 2013 and for follow-up specialty appointments in 2015. Here we evaluated various PHR self-scheduling measures across different EHR departments and estimated operational efficiency, increased revenue, and cost savings attributed to self-scheduling.

Methods
Using scheduling information from our Epic EHR, we examined no show rates, cancellation rates, adoption rates of self-scheduling, and time spent scheduling appointments for primary care (from 2013-2015) and specialty (from 2015) follow-up visits both for self-scheduled and non-self-scheduled appointments, by provider specialty. De-identified Press Ganey patient satisfaction scores for appointments were also examined.

Results
11,807 appointments (90% primary care and 10% specialty appointments) (1.5% of all potentially self-schedulable appointment) were self-scheduled through the PHR from 2013-2015. Self-scheduling rates ranged from 0-9% (by provider specialty). Appointments to dermatologists, optometrist, and gynecologists had the highest self-scheduling rates. Self-scheduled appointments had 14% lower no show rates with primary care providers and 5% lower with specialists; combined no-show rate declined by 9% (p<0.0002 compared to non-self-scheduled appointments). Annualized, at the current decrease in no show rates, over 300 additional appointments, totaling an estimated $15,000 (based on conservative estimate of $50 net revenue for a follow-up visit) in additional net revenue occurred throughout MHS because of self-scheduling. Cancellation rates were higher among patients that self-scheduled (28% in primary care and 30% in specialty care) compared to those that did not self-schedule (13% and 14%, respectively), with p<0.05 for both groups. An estimated average of approximately 4 minutes of staff time (based on time-motion study) were needed to help a patient who does not self-schedule an appointment. Therefore, to date, self-scheduling has saved approximately $16,000 in staff scheduling (estimated staff rates of $20/hr). Almost 1/3 of self-scheduled appointments were made outside of normal business hours (evenings, nights, and weekends) when staff is unavailable. Press Ganey patient satisfaction scores between self-scheduled and non-self-scheduled visit were examined, but there were no statistically significant differences. The rate of self-scheduled appointments doubled annually. We estimate that for each 1% absolute increase in percent of total appointments that are self-scheduled annually that translates into at least $10,000 in increase revenue from decrease in no-show rates and at least $10,000 operational cost savings from enhanced operational efficiency for the MetroHealth System.

Discussion
Real-time, patient visit self-scheduling through a PHR is a growing EHR tethered PHR administrative function. No show rates decreased, probably because of a combination of more favorable initial appointment for the patient and increased ease of cancelling (hence higher cancellation rates) and rescheduling through self-scheduling (cancellations are better than no shows because of ability to reschedule). Also, significant operational staff savings occur because of self-service and better service because of the ability to self-schedule at any time of the day or night.

Conclusion
PHR enabled self-scheduling optimizes use of healthcare resources, both directly by decreasing no show rates and indirectly by saving scheduling staff time and allowing for scheduling outside of regular business hours. Patient appointment self-scheduling can save tens of thousands of dollars even when the percent of appointments self-scheduled is relatively low. As PHR adoption and use increases, appointment self-scheduling could save hundreds of thousands or even millions of dollars annually for large healthcare systems as the majority of follow-up appointments are self-scheduled.
Export data from i2b2 using the new download data web client plugin

Nich Wattanasin, MS¹, Taowei David Wang, PhD¹, Bhaswati Ghosh, MS¹, Reeta Metta¹, Vivian Gainer, MS¹, Shawn N. Murphy, MD, PhD¹²
¹Partners Healthcare, Boston, MA; ²Massachusetts General Hospital, Boston, MA

Abstract

The Informatics for Integrating Biology and the Bedside (i2b2) platform is a framework enabling clinical research and data analysis at more than one-hundred healthcare institutions worldwide. The modularized architecture of i2b2 fosters the rapid creation of plugins to support clinical research needs. At Partners Healthcare, we have developed an i2b2 web client plugin that facilitates the aggregation and exporting of data from i2b2 into CSV format.

Introduction

The i2b2 software has been successfully deployed at more than one-hundred academic medical centers and health care institutions worldwide. The framework provides tools that allow researchers to easily query the data stored in i2b2 at their local site. The i2b2 web client query tool supports Boolean logic-type queries that enables investigators to obtain aggregate counts of patients matching a particular criteria, such as demographics, diagnoses, medications, laboratory values, etc. The modularized design of i2b2 and the fostering of an open source community has led to collaborative projects such as the Shared Health Research Information Network (SHRINE). SHRINE makes it possible to connect any number of i2b2 deployments together to create powerful research networks. Investigators in this instance are able to execute federated queries across multiple participating i2b2 sites to support multi-site clinical research. In both the i2b2 and SHRINE, there is a need for researchers to be able to extract data quickly to continue a necessary patient review process after running an appropriate query.

Methods & Results

At Partners Healthcare, we have developed a web client plugin that enables the exporting of aggregated patient data as a Comma Separated Value (CSV) file. The user first selects a previous query that represents their desired patient cohort. Next, the user selects health-related criteria such as demographics, diagnoses, medications, and/or labs to define the columns of the CSV file. Finally, for each criteria, the user specifies the aggregation method that they desire for that particular column. For instance, an investigator using i2b2 to find suitable patients as part of the clinical trials patient recruitment process would be able to create and download a CSV of patients to review. The investigator could start with a query representing patients with a particular disease. Then, they could choose various medications and lab values, for example, as columns of the CSV output, specifying aggregation preferences such as the existence of a medication, most frequent lab value, or the latest data of a specific disease observation. The resulting CSV output is a single patient per row with the aforementioned specified columns.

Conclusion

The new i2b2 web client plugin has been deployed in Partners Healthcare’s Biobank Portal (Figure 1), an i2b2 implementation bridging clinical research informatics and phenotyping analyses together with the requesting of Biobank samples, such as DNA, plasma, and serum. Researchers using the Biobank Portal use the plugin to obtain a CSV file to review patients that match a criteria to request Biobank samples for. Furthermore, the plugin is intended to help support the facilitation of finding suitable patients for clinical trials across multiple sites using SHRINE as part of the Accrual to Clinical Trials (ACT) project.

Figure 1. Biobank Portal with the new Download Data plugin
Finding Evaluation Frameworks for Complex Health IT Implementations
Jennifer R. Webb, MA, Joshua E. Richardson, PhD, MS, MLIS, Mahima Ashok, PhD, MS, Barry H. Blumenfeld, MD, MS, Jonathan S. Wald, MD, MPH
RTI International, Research Triangle Park, NC

Introduction
In order to plan a program evaluation for a Learning Health System demonstration project at 10 academic centers, we conducted a literature search to identify frameworks to evaluate complex health IT implementations. We developed two research questions: 1) What frameworks for evaluation planning have addressed complex health IT implementations?; and 2) Of those frameworks, what outcomes and measures have been used to assess the success of complex system interventions?

Methods
In partnership with a librarian, we identified multiple National Library of Medicine Medical Subject Heading (MeSH) terms and keywords associated with three high level concepts: evaluation, health IT, and implementations/studies (Table 1). Using 10 different multiple-line search queries using Boolean operators “OR” and “AND” to combine the MeSH terms, we obtained results using PubMed and EMBASE literature databases, and gray literature databases Google Scholar and the New York Academy of Medicine Grey Literature Report Database. Abstracts from the resulting 96 unique citations including two from gray literature were screened for recency (years 2005-2016), English language, human research, adult populations, and relevancy to the two research questions, eliminating 82 citations to arrive at 14 preliminary articles. Following project leader review we abstracted the final set of 7 articles to identify frameworks, their elements, environments where they were applied, and reported evaluation metrics.

Table 1: Example PubMed and Gray Literature MeSH Terms and Keywords

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Health IT</th>
<th>Implementations/Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program evaluation†</td>
<td>Medical Informatics†</td>
<td>Comparative effectiveness research (CER)†</td>
</tr>
<tr>
<td>Evaluation Studies as Topic†</td>
<td>Electronic health records†</td>
<td>Health services research/methods†</td>
</tr>
<tr>
<td>Patient Outcome Assessment†</td>
<td>Health information systems†</td>
<td>“Learning health systems”</td>
</tr>
<tr>
<td>Outcome and Process Assessment (Health Care)†</td>
<td>Medical records systems, Computerized†</td>
<td>Implementation</td>
</tr>
<tr>
<td>Outcome Assessment†</td>
<td>Decision support system, Clinical†</td>
<td>Intervention</td>
</tr>
<tr>
<td>Process Assessment†</td>
<td>Medical Order Entry Systems†</td>
<td>“Patient-centered outcomes research”</td>
</tr>
<tr>
<td>Patient Outcome Assessment†</td>
<td>“Health IT”</td>
<td>“Patient reported outcome measures”</td>
</tr>
<tr>
<td>“Evaluation framework”</td>
<td>“Clinical information systems”</td>
<td>“Patient-generated data”</td>
</tr>
</tbody>
</table>

† = Medical Subject Heading (MeSH) terms

Results
The literature review identified 7 articles that described applied or general frameworks for evaluating complex health IT implementations, including the Consolidated Framework for Implementation Research (CFIR) and Adapted CFIR, the Health IT Reference-based Evaluation Framework (HITREF), the Fit between Individuals, Task, and Technology (FIT) framework, and other unnamed frameworks. The frameworks cover a range of health IT implementation areas including IT adoption, effectiveness of health IT interventions, system development, comparative effectiveness research, and process redesign. Applied settings included a hospital, geriatric center, community health network, and one that addressed multiple settings. Three identified metrics including: end user surveys and time-to-event analyses for clinical alerts. Each framework describes constructs or outcomes for use in a health IT implementation evaluation and lessons learned in using the framework. Although multiple frameworks were identified, some key literature according to experts were not captured in the search including Psek et al. (2015) and Sittig and Singh (2010)†‡.

Discussion
While our detailed search strategy using keywords and MeSH terms identified a number of evaluation frameworks, it also missed some important evaluation framework resources within and outside of published literature. This may reflect poor categorization of concepts and common keywords related to complex health IT implementations. We recommend that researchers supplement search strategies with websites including the EDM forum (http://www.edm-forum.org), CFIR Technical Assistance website (http://cfirguide.org), and the Health Services Research Information Central (HSRIC) website (https://www.nlm.nih.gov/hsrinfo/) that provide useful information and resources for informing evaluation planning of complex health IT implementations.

References
Using Semantic Groupings to Support Clinical Trial Medical Condition Search: A Case Study on Diabetic Complications

Duo (Helen) Wei

Computer Science and Information Systems – BUSN
Stockton University, Galloway, NJ 08205

Abstract
Due to the large volume of clinical trial protocols achieved in ClinicalTrials.gov, it becomes increasingly difficult for patients to find their desirable trials. To expedite the search process, one way is to formulate search queries with standard vocabulary, such as SNOMED or MeSH. Patients can use a medical term and its neighboring terms to formulate their search query keywords. However, this exhaustive search method is time consuming and labor intensive. In this research, we propose to use our previously developed semantic uniformity group (SUG) of SNOMED, which is based on concepts’ properties and hierarchical configuration, to help patients find clinical trials more efficiently.

Introduction
ClinicalTrials.gov is one of the primary sources of information on clinical trials that receives hundreds of thousands of visitors every month. It is routinely used by patients to find clinical trials. The search queries can be grouped into combinations of keywords related to “condition”, “location,” and “treatment” [1]. “Location” and “treatment” are relatively easily to be identified, but “condition” is sometimes very difficult to be described, especially, when patients try to specify a “condition” with a medical terminology with high granularity. One of the options for the patients to formulate their search queries is with the help of the standard vocabulary, such as SNOMED CT. Typically, patients first try to use a medical term in SNOMED CT and its neighboring terms (i.e. its parents, children, siblings) to formulate their search query keywords. However, this exhaustive search method is time consuming and labor intensive, and sometimes, they search in vain since there is a large gap between what they may be looking for and what is available. In this research, we propose a method to use our previously developed semantic uniformity group (SUG) [2] of the SNOMED, which is based on concepts’ properties and hierarchical configuration, to help patients find clinical trials more efficiently.

Method
First, SUGs are generated, which is defined as the maximal set of SNOMED concepts with the exact same set of relationships and sharing the same lowest common ancestor (LCA) within the particular SNOMED IS-A hierarchy. Our assumption is that concepts in the same SUG share same semantics and thus can be used as alternative keywords for retrieval. Second, based on the patients query content, we identify which SUG the concept belongs to. Third, clinical trials are retrieved via the ClinicalTrials.gov API with all members of the SUG as the keywords. Fourth, merge all search results. Fifth, we compare the search results using our method with traditional “exhaustive” search method.

Results
Our method has been applied to the search of “Diabetic complications.” A total of 21 SUGs are generate for concepts with finding site as their relationship for SNOMED (07/2015), where “Diabetic complication” is located in one of the SUGs. There are 13 members in this SUG and we retrieved 1,335 clinical trial protocols with the members as keywords. As a comparison, if we apply “exhaustive” search, then as many as 3,970 protocols are retrieve, which is almost three times if we use SUG members. The search result will be evaluated by patients and domain experts.

Discussion and Conclusion
In conclusion, the SUGs are perceived to be more efficient to support clinical trial search than traditional exhaustive standard vocabulary keyword search, since it significantly condensed search results. We will calculate precision and recall to verify our assumption. Additionally, future research can compare the results using SUGs with that using synonyms-based search method.


1633
A Study of Active Learning for Document Selection in Clinical Named Entity Recognition

Qiang Wei, M.S.¹, Yukun Chen, Ph.D.², Sungrim Moon, Ph.D.¹, Trevor Cohen, M.B.Ch.B., Ph.D.¹, Hua Xu, Ph.D.¹

¹School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA; ²Pieces Technologies Inc., Dallas, TX, USA

Introduction

Effectiveness of machine-learning (ML), a widely used strategy in natural language processing, is largely dependent on the availability of high-quality training data, requiring manual annotations by skilled professionals that can be slow and expensive. It is important to select the most informative sample for annotation to build a high performance ML model. Most of the previous work in active learning (AL) methods apply AL strategies at the sentence level to reduce annotation cost in applying ML methods. Here, we propose AL strategies for named entity recognition at the document level using the i2b2 2010 dataset.

Method

The i2b2 2010 dataset includes 826 annotated clinical notes, of which 477 were used as training set and 349 were the evaluation set. Our program adds a document from unlabeled set to labeled set in each round. The program trains a conditional random field model based on the labeled set and uses the model to evaluate the selection. We used uncertainty-based and diversity-based querying in our study. Uncertainty-based querying uses entropy of entities to measure the uncertainty of labeling the document. Diversity-based querying selects the most representative document in unlabeled set. Random querying and document length-based querying were the baseline.

Result

We used six methods in our study. We generated learning curves of F1 score vs. number of documents (DN) and F1 score vs. number of words (WN) and evaluated our methods. We used area under curve (AUC) as a score for each method (Table 1). Word entropy method had better AUC score for F1 vs. DN than the baseline methods. Normalized word entropy method had the best AUC score of F1 vs. WN amongst all the tested methods.

Table 1. AUC score of learning curve.

<table>
<thead>
<tr>
<th>AUC Score</th>
<th>Random</th>
<th>Document length</th>
<th>Word entropy</th>
<th>Normalized word entropy</th>
<th>Normalized entity entropy</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 vs. DN</td>
<td>0.909</td>
<td>0.919</td>
<td>0.922</td>
<td>0.916</td>
<td>0.905</td>
<td>0.912</td>
</tr>
<tr>
<td>F1 vs. WN</td>
<td>0.908</td>
<td>0.901</td>
<td>0.903</td>
<td>0.910</td>
<td>0.905</td>
<td>0.893</td>
</tr>
</tbody>
</table>

Discussion

In general, the length-based methods performed better on DN as compared to the WN. An explanation may be that longer documents include many informative words, but also have increased number of uninformative words. The entity entropy method considers uncertainty of the entity but ignores the non-entity in a document, causing the lower score. Normalized word entropy method uses the entropies of the entity and the non-entity as well as the length of the document, resulting in the best overall score. The longer document is likely to have a higher Jaccard similarity coefficient, causing similarity-based method to perform worse on WN.

Conclusion

We propose novel active learning methods for named entity recognition task at the document level that can improve machine-learning model performance within limited manual annotation.

References

Using the hierarchical structure of the Medical Subject Headings (MeSH) for automatic MeSH term assignment

Wei Wei, MS, Zhanglong Ji, MS, Lucila Ohno-Machado, MD, PhD
University of California, San Diego, San Diego, CA, USA

Introduction
The hierarchical structure of controlled vocabularies and ontologies has been previously employed in automatic data annotation, such as automatic Medical Subject Headings (MeSH) term assignment. MeSH Now, a MeSH term assignment algorithm developed by the National Library of Medicine (NLM), served as a baseline model for the 2015 BioASQ Challenge (see: http://www.bioasq.org). However, this algorithm does not use the hierarchical structure of MeSH. Our aim was to improve the performance of MeSH Now using this type of information.

Method and Preliminary Results
Given an article, MeSH Now ranks candidate MeSH terms collected from similar articles. Our ranking function considers the contribution of relations (i.e., parents, children, and siblings) to the selection and ranking of potential candidates. For example, if a child of a candidate MeSH term is frequently observed in similar articles, the ranking score of this candidate term is increased accordingly. We followed MeSH Now and used its objective function,

\[ L(Y, G) = - \sum_{i=1}^{n} \Pr(y_i) \log \Pr(g_i) \]  

where \( Y \) is the gold standard, \( y_i = 1 \) if this candidate term is appropriate and \( y_i = 0 \) otherwise; \( G \) is the set of ranking scores; \( i \) indicates the \( i^{th} \) candidate; \( n \) is the number of candidates. In the new ranking function \[2\], a candidate \( x^t \) or one of its related terms, \( x^lr \), is represented by \( m \) features. In the preliminary study, we used two dominant features: (1) the number of similar articles in which this term appears; and (2) the sum of similarity scores of articles in (1).

\[ g_i = \sum_{l=1}^{m} w_l \cdot x^l_i + \lambda \sum_{l=1}^{R} \sum_{i=1}^{m} k_l \cdot x^lr_i \]  

where \( g_i \) is the ranking score of candidate \( x^i \), \( x^l_i \) is the \( l^{th} \) feature; \( w_l \) is the weight of \( x^l_i \); \( x^lr_i \) is the \( l^{th} \) feature of \( x^lr_i \); \( k_l \) is the weight of \( x^lr_i \); \( R \) is the number of relations of \( x^i \); \( \lambda \) is the weight of the contribution of relations.

We explored the contribution of relations using the same training and evaluation data as those used by MeSH Now.

Table 1. Contributions of relations. MAP: Mean Average Precision.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Parent</th>
<th>Child</th>
<th>Sibling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.371</td>
<td>0.371</td>
<td>0.369</td>
<td>0.372</td>
</tr>
<tr>
<td>Recall</td>
<td>0.677</td>
<td>0.677</td>
<td>0.674</td>
<td>0.679</td>
</tr>
<tr>
<td>F1</td>
<td>0.479</td>
<td>0.479</td>
<td>0.477</td>
<td>0.481</td>
</tr>
<tr>
<td>MAP</td>
<td>0.561</td>
<td>0.559</td>
<td>0.559</td>
<td>0.561</td>
</tr>
</tbody>
</table>

Conclusion and Future Work
Siblings of candidates slightly improve the model performance, but not by a significant extent, because the occurrences of parent/child/sibling terms are not frequent enough in similar articles. Perhaps deeper structure information (e.g. edge number) will be necessary improve performance in a significant way. This study was supported in part by bioCADDIE grant NIH/BD2K - U24AI117966.

References
Use of an Interface Terminology for Evaluating Terminology Coverage for a Clinical Decision Support System

Asli Weitkamp, PhD¹, Sina Madani, MD, PhD¹, Trent S. Rosenbloom, MD, MS²
¹Department of Health Informatics Technologies & Services¹;
²Department of Biomedical Informatics, Nashville, TN

Introduction
Clinical terminologies play an important and foundational role in clinical decision support systems (CDSS). Accurate classification of clinical findings is critical for a predictable response from CDSS, while a lack of standardized terminology coverage can become a barrier to deploying and using such systems efficiently. Use of these terminologies can be limited to the manual processes required for identifying and maintaining appropriate linkages to CDSS content. Leveraging interface terminologies designed to represent commonly occurring clinical entities and relationships among them could improve the process of curating CDSS items.

An interface terminology designed to represent clinical problems in a problem knowledge base (PKB) has been in use for a problem list module at the Vanderbilt University Medical Center since 2011. This terminology was designed to contain commonly observed clinical findings, to represent them using general medical terms, and to map to SNOMED CT to support interoperability with other systems. We used existing data in the PKB to identify level of granularities required to represent clinical findings that are used to specify the rules in one of our CDSS modules.

Methods
As part of CDSS rule engine criteria, we identified 12 diseases that, if identified in the outpatient clinical problem lists, resulted in alerts being generated for healthcare providers using the electronic health record systems. We measured PKB usage data, extracted the child concepts of the 12 diseases, and measured their usage in the problem list between 2011 and 2016. We then compared the percentage of usage between the parent and child concepts.

Results
We were able to expand the original list of 12 disease concepts to 175 by including child concepts, based on their usage in PKB (Figure 1). The additional “more granular” child concepts comprised 17% of the total usage of all 175 concepts in the problem list module. The percentage of the child concepts usage, compared to their parent concepts, varied from 0.44% (Coronary Arteriosclerosis) to 100% (Sickling cell disorder) with an average of 25% (Figure 1).

Discussion
SNOMED CT disorder axiom contains clinical concepts that are usually sufficient to be used in clinical decision support applications typically used at the point of care. However, such concepts usually contain more granular child concepts and exposing them all, compared to the actual usage at the point of care, would not only be redundant but also cause a computational explosion of the concepts represented to the providers. On the other hand, based on the CDSS protocols some distinction may need to be made between different subtypes of a given concept (like acute asthma versus chronic asthma). While it can be argued that SNOMED CT has the capability of post coordination in such instances there are still technical difficulties in implementation of this feature in CDSS modules. Therefore, it is important to identify and represent child concepts quickly and objectively for CDSS terminology modules.

Figure 1. Comparison between the usages of 12 disorder concepts together with their children selected for a CDSS module
ItRunsInMyFamily Social Network for Family Health History: Project Update
Brandon M. Welch, MS, PhD,1 Ayesha Aziz, PhD,1 Joshua Schiffman, MD,2
1Medical University of South Carolina, Charleston, SC,
2University of Utah, Salt Lake City, UT

Introduction
Family health history (FHx) is one of the most important tools available for genetic diagnosis and disease risk assessment,1 and is the most consistent risk factor for almost all human diseases. Half of all families display a positive FHx for one or more common chronic diseases.2 Unfortunately, FHx is often inadequately collected in a clinical setting as clinicians are typically hindered by limited time, inadequate reimbursement, and lack of expertise related to genetics and FHx.1 As a result, FHx is discussed with only half of new patients and less than a quarter of return patients.3 As such, a comprehensive FHx data collection and documentation may be best accomplished outside the timeframe of a physician office visit.1 As a result, several researchers and organizations, such as the Surgeon General, have built patient-oriented tools to help individuals collect and record their own FHx. However, these tools tend to be difficult to use, disease specific, collect limited structured data, have limited interoperability, and lack useful clinical decision support. Furthermore, these FHx tools require the patient to recall all their relatives’ health information, which results in inaccuracies and data gaps, particularly in 2nd and 3rd degree relatives. In general, patient-oriented FHx collection tools have had limited adoption and impact outside of research settings.

Approach
To overcome the challenges of current patient-oriented FHx collection tools, clinicians and informaticists at the University of Utah and the Medical University of South Carolina are currently developing an online social network platform called ItRunsInMyFamily. This website will allow individuals to record their own health information and link up with relatives to exchange health information to create a more comprehensive and accurate FHx. ItRunsInMyFamily will leverage both popular social networking techniques and health information technology approaches to help individuals and their relatives collect and share health information with each other. At its core, ItRunsInMyFamily is being developed as a Meaningful Use-aligned, standards-based personal health record (PHR). The platform will use social network techniques to generate a family pedigree of available health information. ItRunsInMyFamily will be a free resource that also provides basic care recommendations (e.g. talk to your doctor about your risk for disease X) to individuals. Clinicians would also be able to access their patients’ FHx information and receive personalized risk assessments and care guidelines for their patients.

Discussion
In this poster, we will describe an update and description of our technical architecture, informatics standards used, the user interface designs, and approaches to data privacy & security being incorporated in the development of ItRunsInMyFamily. A beta version of the site is expected to be available for AMIA 2016.

Conclusion
We believe that an online platform using social networking technology which helps individuals share and discuss their health information with relatives will create a more comprehensive and accurate FHx than current approaches.

References
Improving the Workflow of Curbside Consultation by Using Unstructured Clinical Notes - a Natural Language and Machine Learning-based Approach

Wei-Hung Weng, MD¹,², Avni M. Khatri, BS², Kavishwar B. Wagholikar, MBBS, PhD²,³, Adam B. Cohen, MD⁴, Henry C. Chueh, MD²,³
¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA; ²Laboratory of Computer Science, ³Department of Medicine, ⁴Department of Neurology, Massachusetts General Hospital, Boston, MA.

While caring for patients, physicians often encounter clinical questions beyond their expertise. To resolve these questions, physicians refer the patient to a relevant specialist. But often the physicians tend to reach out to the specialist themselves using phone and email. Such informal discussions are defined as curbside consultations. It is challenging, however, to find a relevant specialist for curbside consultation. Consequently, the curbside consultations can be inappropriate, time-consuming, and even unsafe for patients.

To resolve these problems, expert recommendation systems (ERS) have been developed to identify suitable specialists, who can appropriately answer the clinical problem. Yet ERS have not been widely used clinically due to lack of performance. As such, natural language processing (NLP) and machine learning (ML) have been shown to improve the validity of ERS by identifying biomedical experts. In this study, our objective is to develop an NLP and ML-based clinical ERS.

We hypothesize that an NLP-ML model trained on clinical notes mapped to specialties and authors can be repurposed as an accurate ERS. We will develop NLP-ML software and model based on clinical notes from Massachusetts General Hospital that can predict the specialty and the author of a given note. To evaluate the performance of the model for the ERS task of identifying the appropriate specialty and specialist given a clinical question, we will then construct a gold standard by interviewing the clinical specialists. The participating specialists will be requested to provide questions, which have been consulted in curbside consultations. The specialists will also provide the name of the domain experts that appropriately resolved the question. For evaluation, the questions provided by the specialists will be input into the NLP-ML model, and the output will be compared to the expert-list provided by the specialists. Figure 1 gives the framework of the study.

UMLS concepts in the clinical notes will be extracted using NLP tools. The concepts along with the specialties or author names will be fed into supervised ML algorithm for modeling, and the model performance can be evaluated using normalized discounted cumulative gain (NDCG). To apply the model to an ERS task, we will use the NLP tool to extract UMLS concepts from the gold standard questions. The concepts will be input into the model to obtain a ranked list of specialists. The list will be compared against the gold standard and evaluated by NDCG. We will explore different feature extraction and ML techniques to optimize the model. For preliminary work, we have developed an NLP-ML pipeline for predicting the medical domain from a clinical note, using 431 publicly available notes in iDASH repository. Cross-validation revealed that NDCG improved from 0.905 with baseline bag-of-words to 0.985 with UMLS concepts and maximum entropy algorithm (table 1). We adopted UMLS concepts to reduce dimensionality and capture semantically related concepts inside the content. Linear SVM, random forest, and maximum entropy algorithms were selected for effective NLP tasks.

The NLP-ML approach may enhance the real-time ERS, and ultimately optimize the workflow of curbside consultation, save time for physicians and improve patient care.

---

Table 1. NDCG of ML models

<table>
<thead>
<tr>
<th>Model Description</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-words + vector similarity</td>
<td>0.905</td>
</tr>
<tr>
<td>UMLS + vector similarity</td>
<td>0.901</td>
</tr>
<tr>
<td>UMLS + linear SVM</td>
<td>0.926</td>
</tr>
<tr>
<td>UMLS + random forest</td>
<td>0.976</td>
</tr>
<tr>
<td>UMLS + MaxEnt</td>
<td>0.985</td>
</tr>
</tbody>
</table>

Figure 1. Study design
Patient Preferences for Visualizing Possible Outcomes of Breast Reconstruction

June Weston, BS¹, Mary Catherine Bordes, BS¹, Sheng-Cheng Huang, PhD¹,², Gregory P. Reece, MD¹, Fatima A. Merchant, PhD³, Scott B. Cantor, PhD¹, Mia K. Markey, PhD¹,², Michelle C. Fingeret, PhD¹

¹The University of Texas MD Anderson Cancer Center, Houston, TX;
²The University of Texas at Austin, Austin, TX;
³University of Houston, Houston, TX

Abstract
The long-term goal of our research is to enhance the consultation process for women undergoing breast reconstruction surgery. The purpose of the current study was to conduct focus groups to evaluate patient experiences related to the breast consultation process and to obtain feedback on components of our prototype decision support system.

Purpose
The long-term goal of our research is to enhance the consultation process for women undergoing breast reconstruction surgery. Our vision is a decision support system that will enable breast cancer patients, in consultation with their healthcare providers, to choose a reconstruction strategy with maximal potential to optimize psychosocial adjustment. This system for shared decision-making will use quantitative models to tailor the presentation of patient-specific information about breast reconstruction outcomes. The purpose of the current study was to conduct focus groups to evaluate patient experiences related to the breast consultation process and to obtain feedback on components of our prototype decision support system.

Methodology
The study sample was drawn from patients seeking treatment at The University of Texas MD Anderson Cancer Center and from local breast cancer support groups in the surrounding Houston area. A trained facilitator used a semi-structured interview guide to conduct each focus group (Figure 1). Notes and audio recordings of the group discussion were transcribed and analyzed with two basic approaches: qualitative/ethnographic summary and a systematic coding via content analysis.

Results
Twenty-one women participated in one of six focus group. Each focus group contained between 2-5 participants. Several key themes related to visualization emerged from the focus groups: (1) Women mostly preferred to see their own image rather than the image of someone else; (2) 2D images were described as more natural looking and strongly preferred by all participants when compared to 3D still images; (3) 3D still images were described negatively in all groups: used terms like “creepy” “weird” “bionic” “Halloween;” (4) Smooth and mesh textures were poorly received; (5) Video and interactive 3D photo were very well received by all groups.

Conclusions
Interactive photos and videos appear to show particular promise for helping patients to understand possible appearance outcomes from breast reconstruction.
Getting to the Right ICD-10 Code: A Multifaceted Approach

Nancy Whipple, BSN, MS, MM\(^1\), Harley Ramelson, MD, MPH\(^{1,2,3}\)

\(^1\)Partners HealthCare System Inc, Wellesley MA; \(^2\)Harvard Medical School, Boston, MA; \(^3\)Brigham and Women’s Hospital, Boston, MA

The transition from the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) to ICD-10-CM coding system for billing claims codes occurred on October 1, 2015. Several components essential to the successful transition to ICD10 included adequate education and training, proper selection of software vendors and tools, exhaustive internal and external quality assurance testing, and proper change management processes and communications. This poster describes the changes we made to our EHR software system to support our clinicians in selecting an appropriate and specific diagnostic ICD-10 code for ambulatory practices.

One of the primary goals of our ICD-10 project was to optimize the process of selecting an appropriate ICD-10 diagnosis code by both presenting pertinent clinical questions important to the selection of a specific ICD-10 code and minimizing the number of mouse-clicks and screen changes. Based on a user-centered iterative design feedback process, we offered the ability to select an ICD-10 code in a variety of ways: Search, Favorites, Templates, Past Patient Data, and Past Encounter Data.

**Search:** Users can enter one or more text terms and be presented with a set of matching results tailored\(^1\), when appropriate, to a patient’s age, sex and medical history. These results contain root codes and/or specific billable ICD-10 codes. A root code is one that encompasses a group of related ICD-10 codes, is not billable, and requires further refinement to greater specificity to get to a billable code. From this result list, if a user selects a billable ICD-10 code, that code is added immediately to the encounter; if a user selects a root code, additional questions are presented. Basically, each question represents an additional character in the refined code. The questions are based on the National Library of Medicine’s mapping rules to ICD-10 and were enhanced and customized to cover a greater number of ICD-10 codes and to be more efficient from a clinical standpoint. Due to dependencies among questions, branching logic was instituted so that an answer to one question would affect the subsequent questions and answer sets presented to the user. The question screen is interactive so that if the clinician changes a response to any question, the screen reflects the updated valid questions and answers.

**Favorites:** The user or practice can create a list of commonly used diagnosis codes and access these at the time of documenting billing codes. **Templates:** Templates are collections of a variety of billing codes and information important for actions performed at the end of a visit. These include diagnosis codes, procedure codes, orders, E&M codes, follow-up information, referrals and educational materials to provide to the patient. Templates can be tailored to specific visit types (e.g. a 6 month well baby visit) or conditions (e.g. diabetes follow up visit), or can be based on age/sex of a patient (e.g. male adult senior annual exam). **Past Patient Data:** Users can use the patient’s current list of problems to get to a matching ICD-10 code or set of codes. This is accomplished using maps between SNOMED-CT and one or more corresponding ICD-10 codes. In addition, ICD-9 diagnosis codes added prior to Oct. 1, 2015, can be selected and a corresponding ICD-10 code will be added to the current encounter. Subsequent to this date, a reverse chronological list of previously selected ICD-10 codes for that patient can be accessed and one or more codes can be selected and added to the current encounter. **Past Encounter Data:** Diagnoses and individual elements of previous encounters can be selected and “copied forward” to the current encounter. The clinician can also copy the entire previous encounter in a single step.

Data from Oct 2015 to March 2016 showed that clinicians were employing many of the tools we provided. There was a significant increase in the number of favorite lists created. We also found a high correlation between the codes in templates and in favorite lists and codes entered in encounters.

**References**

Modifying a Case Report Form to Collect Anonymous and Consistent Data

Elizabeth White\textsuperscript{1}, Sarah Collins RN, PhD\textsuperscript{1,2}, Beatriz H. Rocha MD, PhD\textsuperscript{1,2}, Neelima Karipineni MD, MMSc\textsuperscript{1,2}, Taylor Read MS\textsuperscript{1}, Saverio Maviglia MD, MS\textsuperscript{1,2}, J Perren Cobb MD\textsuperscript{3}, Roberto A. Rocha MD, PhD\textsuperscript{1,2}

\textsuperscript{1}Brigham and Women’s Hospital, Boston, MA; \textsuperscript{2}Harvard Medical School, Boston, MA; \textsuperscript{3}University of Southern California, Los Angeles, CA

Introduction

In multi-site studies involving data transfer, the security of patients’ protected health information (PHI) is vitally important. Many strategies exist to de-identify study data \cite{1}, but when patient identifiers are not necessary to achieve study aims, it is more efficient to refrain from collecting PHI. The objective of this project was to implement an existing case report form (CRF) with specific modifications to remove PHI and preserve data integrity. These modifications should streamline data transfers while maintaining the ability to answer key research questions.

Methods

We used a validated paper CRF\textsuperscript{1} and implemented it in electronic form using Research Electronic Data Capture (REDCap). REDCap is a secure, web based data collection application \cite{2}. After implementing the electronic CRF (eCRF), data fields were reviewed to determine which ones constituted PHI and which ones required modification to promote data consistency. The identified fields were deleted or modified to address these needs, and in the case of data consistency, several fields were also added. The resulting eCRF was tested and validated by a team of informatics and clinical subject matter experts (SMEs).

Results

To remove PHI from the eCRF, 10 changes were made. Seven of the changed fields were dates, of which 3 were deleted and 4 were modified to a time reference format rather than an exact date. As a result of the changes to date format, 2 fields became unnecessary and were deleted. Patient medical record number (MRN) was removed from the eCRF and replaced with an automatically generated REDCap identification number. To improve data reliability, 27 additional changes were made to the eCRF. Nine fields were added to prevent data loss or resolve ambiguities. After these changes, 1 field became redundant and was deleted. Finally, 17 fields were modified in either the field label or answer choices in order to provide more clarity to the user regarding what data was being asked for and what units were required. Seven of these modifications were to replace the answer choice “NA,” which our SMEs determined to be ambiguous, with “Data not available.” An example of each type of change is shown in Table 1.

\begin{table}
\centering
\begin{tabular}{|l|l|l|l|l|}
\hline
Original Field Label & Original Answer Format & Modified Field Label & Modified Answer Format & Reason for Modification \\
\hline
Symptom onset date of first/earliest symptom & DD-MM-YYYY & Onset of first/earliest symptom & 1, 0-1 day before admission | 2, 2-3 days before admission | 3, 4-5 days before admission | 4, 5 days or more before admission | 999, unknown & Remove PHI \\
\hline
Other & 2, YES - Confirmed | 1, YES - Probable | 0, NO & Other Infectious Respiratory Diagnosis & [not modified] & Improve data reliability \\
\hline
\end{tabular}
\caption{Examples of changes to eCRF to remove PHI and improve data reliability}
\end{table}

Conclusion

Our objectives were met through this process, which took approximately 20 hours of total effort. With these modifications, participating sites have obtained expedited IRB approval and data collection is about to begin. We will monitor the effort required to complete the eCRF, as well as suggestions for improvement. Once the study is concluded, we plan to compare the data obtained with this revised eCRF with data from sites using the original CRF.

References


\textsuperscript{1}RAPID CASE RECORD FORM - Severe Acute Respiratory Infection: https://isaric.tghn.org/site_media/media/articles/ISARIC-WHO-SARI-BSP_Master_Case_Record_Form_29JAN14.pdf
Towards A Better Cancer Care Through The Use Of A Chemotherapy Ordering Information System: Transferable Lessons Learnt

Nilmini Wickramasinghe 1,3, Peter Haddad 1,3, Stephen Vaughan3, Jonathan L. Schaffer2

1Deakin University, Melbourne, Victoria, Australia, USA, 2 Cleveland Clinic, Cleveland, Ohio, 3Epworth HealthCare, Melbourne, Victoria, Australia

Abstract

Today, cancer is one of the most prevalent diseases. Cancer treatment regimens are complex, are associated with high risks and often have unwanted side effects. The outcomes of a longitudinal study to evaluate the impact of using an American computerized physician order entry (CPOE) system to improve cancer care outcomes in an Australian not-for-profit tertiary healthcare environment are presented. Results indicate that while the system is necessary it is not sufficient to provide superior care. Finally, lessons learned are presented.

Introduction

Cancer is one of the most common chronic diseases in today’s modern society. The treatment regimens used to treat cancers are complex, associated with high risks and often have unwanted side effects and complications. Within cancer services, the administration of chemotherapy is typically a major component in the treatment plan. The use of a computerised physician order entry (CPOE) system as a chemotherapy ordering system (COS) for the planning, ordering, dispensing and administration of these cytotoxic drugs should improve the quality, safety and efficiency of this process. A major consideration is the effectiveness of the implementation of such a system, the facilitators and barriers to this process, as well as how to address them. These points form the central focus of this study.

The Proposed Model

Actor Network Theory [1] and Agency Theory [2] were combined to provide the theoretical lens and guide the systematic literature review to explore the key points necessary to successfully implement the COS (chemotherapy ordering system) in the Australian not-for-profit tertiary healthcare environment. Combining the results from the preceding literature synthesis with an initial assessment of the data, the conceptual model was developed (Figure 1).

Methods and Materials

Methods: The study adopts a multi method approach that is predominately qualitative using semi-structured interviews, focus groups and survey techniques to gather critical information.

Materials: An online survey was designed, developed and then validated to assess usability and acceptance of the COS and data triangulation was adopted to ensure internal and external validity. Specifically, the survey targeted three categories of healthcare professionals and executives as follows: Oncologists, Executives and Casual Users and COS Daily Users.

Results:
The COS was found to be necessary for many elements of care and administration, i.e.; EMR elements available to track clinical documentation in concert with ordering and results documentation, but not sufficient to provide superior care. A key challenge relating to poor goal aligned behavior emerged. Agency Theory underscores the importance of engaging users to ensure strong goal alignment between their tasks and their interactions with the system and the hospital's goals and objectives [13], yet this was poorly addressed in the specific context. Key lessons include the need to engage with all users, to appropriately design training as well as effecting thoughtful change management and organization culture/subculture modification changes.

Conclusions:
The adoption and implementation of COS requires a holistic perspective during each phase of design and implementation. Adequate attention to the health ecosystem, the system structure, the delivery operations and the satisfaction of all the respective users e.g., oncologists, nurses, pharmacists, GPs and administrators as well as the nested clinical structures should be addressed.

References

E-prescriptions and Problem Lists: Looking for Indications Using the Open-Source MEDI Medication-Indication Matching Resource

Taylor Woodroof, PharmD; Wing Liu, PharmD; Scott D. Nelson, PharmD, MS

1Vanderbilt University Medical Center, Nashville, TN

Introduction
When performing medication reconciliation and medication therapy management, it is critical that prescribed medications are reconciled with indications on the patient’s problem list and that untreated diagnoses are re-evaluated. However, these reconciliation processes are dependent on the completeness of the medication and problem lists, and could be aided by automated matching. The purpose of this study was to estimate the proportion of e-prescribed medications at our facility with a potential indication listed on the patient’s problem list using the open-source medication-indication matching resource MEDI (MEDication Indication) to infer probable matches.

Methods
We analyzed a de-identified dataset of medications e-prescribed from 1/1/2015 to 6/30/2015 at Vanderbilt University Medical Center (VUMC). To be included in the study, patients have to have established care at VUMC, defined as ≥2 visits in the past year and ≥1 e-prescribed medication. We matched e-prescriptions to potential indications on the problem list using MEDI (MEDI_01212013), which integrates medication indication information from four public medication resources, RxNorm, Side Effect Resource (SIDER) 2, MedlinePlus, and Wikipedia.1 To improve precision of the medication-indication pairs, we separately analyzed the MEDI High Precision Subset (HPS) (MEDI_01212013_HPS), which is a smaller set of medication-indication pairs found within RxNorm, or at least two of the other three resources. We mapped prescription medication codes from First Databank to RxNorm generic drug ingredient names (RxCUI, TTY=IN). Prescriptions were excluded if they consisted of multiple ingredients or did not map to RxNorm. In order to match diagnosis on the problem list to the MEDI dataset, we mapped SNOMED CT problems to ICD-9 codes. We established a gold-standard to evaluate the MEDI precision and recall by having two pharmacists independently manually review e-prescriptions and problem lists from 30 randomly selected patients. Disagreements were resolved by discussion and consensus. For analysis, we calculated precision, recall, and descriptive statistics.

Results
Out of 62,191 patients included in the study, there were 270,045 electronic prescriptions sent. 61.3% of electronic prescriptions had an indication match within the patient problem lists, whereas only 37.5% had a match using MEDI-HPS. Using MEDI, patients had an average of 1.3 medications match to prescribed indications and 0.44 medications using MEDI-HPS. For comparison, the final dataset contained an average of 4.3 prescribed medications per patient. For the gold-standard comparison, the reviewers had an inter-rater reliability of 0.81. The precision of MEDI compared to the gold standard was 55.8% with a recall of 61.4%; whereas MEDI-HPS had a precision of 85.2% and recall of 95.8%. The top-20 most frequently e-prescribed medications had an indication match 61% (n=47,892) of the time using MEDI, with only 36.8% (n=28,735) having an indication using MEDI-HPS. Secondary analysis excluding medication prescribed with a day supply <14 days gave slightly better, yet similar results (data not shown).

Conclusions
We were not able to match indications from the patient problem list for the majority of e-prescribed medications with an acceptable level of precision (MEDI-HPS). Manual review showed that MEDI-HPS has a higher precision and recall than MEDI. This marks the first use of MEDI-HPS matching e-prescribed medications with indications. Given its high precision and recall after manual review, this study suggests that many problem lists (and potentially medication lists) are incomplete in the EHR, emphasizing the difficulty, and potential beneficial use cases, of using an automated medication-indication matching resource such as MEDI-HPS.

References
Effect of Alcohol Prohibition on Drug-related Mortality

James A. Woodward, B.A.¹, Ray H. Hylock, PhD¹, Qiang Wu, PhD¹, Xiaoming Zeng, PhD¹
¹East Carolina University, Greenville, NC

Overview

Since 2000, overall mortality rates in the United States have decreased 16.6%¹ (Figure 1) while drug-related mortality rates have increased 137%². This research investigates how the county alcohol prohibition status (i.e. whether alcohol sales are completely prohibited (dry counties), completely permitted (wet counties), or there is some regulation of alcohol sales (mixed counties)) is related to increased drug related mortality rates. We performed a policy level literature search that revealed evidence of a positive relationship between alcohol prohibition and meth lab seizures³, but found no research directly related to alcohol prohibition and drug-related mortality.


Methods

County and State name, county FIPS code and alcohol prohibition status were parsed from the national alcohol prohibition map SVG file. The wet-dry status, Census data, and CDC mortality datasets were created, normalized and merged in SAS, which was also used for descriptive and analytical statistical analyses.

Results

Our analysis reveals that dry counties consistently have higher mortality rates than both mixed and wet counties through all years of the analysis (2002-2014) and that dry counties have statistically higher odds of having high drug mortality rates (>20 drug related mortality rate) than wet or mixed counties (Figure 2).

References

https://www.census.gov/popest/data/counties/asrh/2015/index.html
Restoring line breaks in Epic-derived clinical notes

Stephen T. Wu, PhD1,2, Allison Sliter1, Meikun Wang1, Tamara Timmons, MD1, Steven Bedrick, PhD1

1Oregon Health & Science University, Portland, OR; 2Mayo Clinic, Rochester, MN

Introduction

Secondary use of the Electronic Health Records (EHRs) is challenging for multiple reasons. Unfortunately, with such vast amounts of heterogeneous data, small design decisions in the implementation of a system can have wide-ranging impact on the viability of the data for downstream processes.

In the early stages of a patient-level information retrieval (IR) and natural language processing (NLP) project at Oregon Health & Science University, we discovered that data originating in free-text note fields had been modified, and all line-ending characters had been replaced with a variable number of whitespaces. This profoundly affects a number of downstream NLP tasks, such as section detection, as well as any analysis of semi-structured data such as lists, tables, etc. that may have been present in the notes. Further investigation of these errors pointed back to the Extract-Transform-Load (ETL) processes within the Epic EHR that produces the standard Clarity database; thus, this transformation of line endings is a barrier to local IR and NLP research, and also to secondary use of clinical notes in any institution using the same Epic ETL process.

Methods

Inspired by previous approaches to sentence detection (1), we developed a conditional random field (CRF) model (2), using the CRF++ package, to restore line breaks on data extracted from Epic's Clarity. For each word \(w_i\) in clinical text, we sought to predict whether the following delineation should be a <space> or a <new line>.

Features for the CRF classifier were developed on a small development set of "synthetic" line-break data. Namely, we took a de-identified data set (from a non-Epic EHR) with original line breaks preserved, then removed line breaks. We could then compare the original line breaks with those hypothesized by our classifier.

In this development data, we noted that the appearance of line breaks was highly correlated with other textual features in close proximity. Our basic feature set included case information, the presence or absence of punctuation medially or token-finally, and whether or not the token contained digits. In addition to this, we included secondary features that were combinations of these features, and calculated these features in 2-to-6-word contexts around a candidate line ending location. We built a classifier to predict which white space characters between tokens might have originally been a newline character. We found that 4-gram contexts were the most successful.

Evaluation and Deployment

For evaluations, we annotated 450 line-break-less Epic clinical notes, introducing 11,140 line breaks. We performed cross-validation tests on this data, resulting in scores of 96.5%, 87.5%, and 91.7% for Precision, Recall, and F1-score. This performance on real data was lower than on synthetic development set, but it is still an important improvement over the existing lack of line break data in Epic-derived clinical text.

We parallelized the line break-restorer and ran it on 10,111,930 clinical notes to be used in a new IR collection (3). Future work includes improved feature extraction, a characterization of performance on different types of EHR text (e.g., Pathology vs. Radiology reports), and making our line-break restorer available open-source to other institutions with similar issues arising from their Clarity databases.

References

What Can Neural Networks Learn from Unlabeled Clinical Narratives?
Yonghui Wu, PhD1, Jun Xu, PhD1, Yaoyun Zhang, PhD1, Hua Xu, PhD1
1School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction: Word embeddings can capture abundant semantic relations to extend the current applications of natural language processing (NLP). Researchers have applied word embeddings in the clinical domain and reported performance improvements in many clinical NLP tasks. However, there is no research to examine the details of semantic relations captured from clinical narratives, as it's a non-trivial and challenging task in general. A good interpretation of the "black box" (word embeddings) can help people better understanding why neural word embeddings could help and what is captured in the real-valued numbers. This study attempts to interpret the details of semantic relations captured from large clinical narratives using neural networks. We looked into word embeddings trained from a 2.2 G clinical narrative corpus and examined the semantic relations. Our results showed that in addition to the conventional semantic relations defined by linguistics, such as the word prefix/suffix, part of speech tags (POS), the word embeddings also captured more complex semantic relations (medications, body locations, and even complex logical relations, such as $WED<$"father"$, "$mother"> \approx WED<$"wife"$, "$husband">$, where "WED" denotes "word embedding distance", which is defined as the cosine-similarity score between two word vectors).

Methods: In our previous work1, we trained word embeddings from 2.2 G unlabeled clinical narratives from MIMIC II corpus using a neural network. The training corpus was composed of various types of notes, including discharge, radiology, ECG, and ECHO notes. All the words with frequency great or equal to 5 (about 41,000 words) were used. We applied t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to calculate the neighbors and generate word cloud visualization, where words formed various clusters according to the distance to its neighbors. Then, we examine the word clusters and tried to label the semantic relations among the words within each cluster. For example, the label “plural/singular” was assigned to the cluster containing the plural form and singular form of a word. Finally, we compared our results with the semantic relations reported in the previous research to compare the differences and report new findings.

Results: Although clinical narratives are often criticized as ungrammatical and telegraphic, we found that the neural networks captured many reasonable semantic relations. Table 1 shows some examples of captured semantic relations. As shown in Table 1, the word embeddings captured not only the relations defined by linguistics, such as words with the same prefix/suffix, POS, but also medical and complex relations. Figure 1 shows the visualization of a complex relation: $WED<$"father"$, "$mother"> \approx WED<$"wife"$, "$husband">$, similar to the <"king"$, "$queen"> and <"man"$, "$woman"> relation reported in the general English domain.

<table>
<thead>
<tr>
<th>Captured relations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plural, singular</td>
<td>pt/pts, lobe/lobes, segment/segments</td>
</tr>
<tr>
<td>Suffix</td>
<td>hematology, neurology, cardiology</td>
</tr>
<tr>
<td>POS</td>
<td>take, start, continue, check, monitor</td>
</tr>
<tr>
<td>Tense</td>
<td>measured, recorded, observed, evaluated</td>
</tr>
<tr>
<td>Medications</td>
<td>insulin, lasix, metoprolol, lisinopril</td>
</tr>
<tr>
<td>Body locations</td>
<td>ankle, wrist, elbow, knee, shoulder</td>
</tr>
<tr>
<td>Modifier</td>
<td>minimally, severely, partially, somewhat</td>
</tr>
<tr>
<td>Disorder</td>
<td>enlarged, thickened, dilated, ruptured</td>
</tr>
<tr>
<td>Complex relations</td>
<td>$WED&lt;$&quot;father&quot;$, &quot;$mother&quot;&gt; \approx WED&lt;$&quot;wife&quot;$, &quot;$husband&quot;&gt;</td>
</tr>
</tbody>
</table>
Diabetes Self-Management Applications: Focus Group Findings from Elderly Diabetic Patients

Qing Ye, MS\textsuperscript{1}, Suzanne A. Boren, MHA, PhD\textsuperscript{1,2}, Uzma Khan, MD\textsuperscript{3}, Min Soon Kim, PhD\textsuperscript{1,2}

\textsuperscript{1}Informatics Institute; \textsuperscript{2}Department of Health Management and Informatics; \textsuperscript{3}Department of Medicine, University of Missouri, Columbia, MO

Abstract

We conducted two focus groups with 10 older diabetic patients to identify barriers in functionality and usability of mobile diabetes self-management apps. We collected demographics and diabetes self-management app experiences. They completed a set of diabetes management tasks using nine representative diabetes self-management apps on iPad. Descriptive statistics and thematic analysis were used. We found current diabetes apps are not usable and do not fulfill the needs of older diabetic patients for self-management.

Introduction

The prevalence of diabetes is increasing, and 25% of Americans above 65 years old have diabetes. The number of diabetes self-management (DSM) applications available has risen. However, we do not know whether these apps provide effective DSM for elderly diabetic patients. The purpose of this study was to identify barriers in usability and functionality related to the needs of elderly diabetic patients for DSM apps.

Methods

We conducted two focus groups with 10 older diabetic patients. After the multi-step review process, 73 iOS apps were identified eligible. We selected apps on the basis of app ratings, number of reviews, and presence of seven principles by the American Association of Diabetes Educators (AADE7) Self-Care Behaviors\textsuperscript{TM}. The seven behaviors are Healthy Eating, Being Active, Monitoring, Taking Medication, Problem Solving, Reducing Risks, and Healthy Coping. Nine apps were selected and installed on iPads: MySugr Diabetes Logbook, GoMealsHD, DiabetesConnect, Diabetes Pilot HD, Tactio Health, Diabetes App Lite, ezbdz, Daily Carb Premium, Diabetes in Check. We collected demographics, diabetes history, Single item literacy screener question (SILS, 1-5), and prior experience using smart mobile devices and DSM apps. SILS scores greater than 2 are considered positive, indicating some difficulty with reading health related material. The participants were asked to perform self-management tasks related to Healthy Eating, Problem Solving, and Healthy Coping, which are the critical principles for improving self-care behaviors. We collected their preferences, issues, and needs of DSM app features. System Usability Scale (SUS, 0-100) was used to measure usability. Descriptive statistics and thematic analysis were used to identify barriers elder patients encountered as they interacted with the DSM apps.

Results

The average age of participants was 67 years old. Of the 10 participants, six were male. All participants were white. Six of them were Type 1 diabetes and four were type 2 diabetes. The average diabetes history was 27 years. Most participants had high health literacy (SILS=1.8). More than half of the participants had used smart mobile devices for over one year, however nearly all participants had never used diabetes apps to help them manage their condition. The features that participants liked most for their DSM were documentation, information, and goal setting. Most participants thought these apps helped them gain the skills of counting carbohydrates (72%), setting goals for a healthy lifestyle (62%), and reading food labels (51%) in Healthy Eating, and recognizing and reacting to high or low blood sugar levels (70%) in Problem Solving. On the other hand, participants thought these apps did not help gain the skills of motivate exercise/meditate (16%) and pursue hobbies (32%) to deal with stress, anxiety or depression in Healthy Coping. Thematic analysis revealed that usability was their primary concern about DSM apps in managing diabetes conditions (i.e. complicated layout, not easy to use, and not useful). The average score of the SUS was 49 (SD=28), which was considered not acceptable. Besides the usability issue, participants considered presence of documentation functionality and cost as important factors when they decided to download diabetes apps.

Conclusion

Older diabetes patients would like to choose apps that are easy to use, allow documentation of glucose levels, diet and physical activity. This study suggests current DSM apps do not provide meaningful features for self-management and may not fulfill the needs of elderly diabetic patients. Future diabetes apps designed for diabetic users should be easy to use for older patients also.
A Review of Mobile Phone-based Interventions and Applications for Medication Adherence

Po-Yin Yen, RN, PhD, Jessica M. Garvey Smith, BS, Michelle P. Zhou, Megan Chamberlain, Xiaonan Ji, MS, Albert M. Lai, PhD
Department of Biomedical Informatics, The Ohio State University, Columbus, OH

Abstract
Medication non-adherence is a critical problem, resulting in poor clinical outcomes and wasteful spending. We reviewed 803 articles in PubMed and found 80 relevant studies with mobile phone interventions. We also reviewed 10 free mobile applications with the highest downloads and user ratings. We identified limitations and gaps for future research.

Introduction
In 2010, $100-300 billion of US spending on healthcare costs was due to medication non-adherence, leading to re-hospitalizations, mortality, and multiple morbidity.\(^1,2\) Portable technology such as smartphones has the potential to improve medication adherence with reminders and educational messages.\(^3\) The purpose of the study was to review the effectiveness of such an mobile phone intervention and identify gaps for future research.

Methods
We searched in PubMed with the following search strategy: ((Medication Adherence[MeSH Terms] OR "medication adherence" OR "medication compliance") AND (technology OR application OR "mobile phone" OR smartphone). The inclusion criteria was that studies must utilize mobile phones as an intervention. We also searched in Google Play for free medication adherence applications with the highest number of downloads and at least 4-star user ratings. We investigated study design, trial length, study population, outcomes, and mobile phone features used for medication adherence.

Results
A total of 803 articles were found in PubMed; Only 80 studies were included with a mobile phone intervention for medication adherence. Of the 80 studies, 66 (82.5%) solely used text/voice messages or personal phone call as reminders and educational messages; only 14 (17.5%) used smartphone applications with advanced features, such as adherence monitoring and health status tracking (e.g., blood pressure, glucose, weight). The most common study populations are HIV, followed by diabetes and general non-specific population. Most studies conducted randomized controlled trials with less than 6 months intervention; Only 12 studies followed up with study subjects for 1-2 years. 42 (52.5%) found the mobile phone-based intervention effective, 13 (16.2%) found the intervention inconclusive, 10 (12.5%) found no improvement, and 15 (18.8%) with pending results. In addition, we reviewed 10 free mobile applications. The most common features included reminders, management of dependents/multiple accounts, inventory tracking, medication adherence on-time records, adherence tracking, important contact information, usage of medication pictures or icons, and health conduction tracking. We did not find any research regarding the effectiveness or usability of these applications.

Conclusion
Patients with chronic conditions are likely to stop taking their medication within the first year.\(^2\) Medication adherence research should aim to conduct studies for more than one year. Smartphone applications with advanced features were perceived useful, but lacked long-term and effectiveness studies. A medication adherence intervention could be simple text messages or complex smartphone applications, economic or costly monitoring systems (e.g., electronic pill trays/bottles). In order to benefit the majority of patients, there is a need to investigate cost-effective solutions. Further research should include long-term adherence trials, usability and feature comparative studies, and application development for complex and multiple medication regimens.

References
Vision-Based Hand Hygiene Monitoring in Hospitals

Serena Yeung, MS1, Alexandre Alahi, PhD1, Albert Haque, BS1, Boya Peng, BS1, Zelun Luo, BS1, Amit Singh, MD2, Terry Platchek, MD2, Arnold Milstein, MD1, Li Fei-Fei, PhD1

1Stanford University, Stanford, CA; 2Lucile Packard Children's Hospital Stanford, Stanford, CA

Introduction

Recently, much progress has been made in machine learning-based interpretation of clinical data for decision support and knowledge discovery. These works have reasoned on data such as electronic medical records and radiographic images. However, a valuable source of clinical data that is still underexplored is visual data capturing patient experience and environment during health care episodes such as hospital stays. Such data can contain rich information about patient condition, and the occurrence and characteristics of clinical care activities.

In this work, we interpret visual clinical data for the application of health care-associated infection (HCAI) prevention, which is a challenging and costly problem for hospitals in the United States1. Proper hand hygiene is known to play an important role in preventing HCAI2. We introduce an approach for monitoring hand hygiene compliance using machine learning-based interpretation of visual recording of the environment. Specifically, we propose to deploy privacy-safe depth sensors in hospitals to capture the physical space near hand hygiene dispensers, and use computer vision methods to detect hand hygiene actions.

Method and Experiments

We use a convolutional neural network3 to detect whether hand hygiene occurs in a video frame. The network consists of 2 convolutional layers and 2 fully connected layers, with a binary classification output, and we optimize it using stochastic gradient descent. We collected a pilot dataset of 20 hours of depth signals (Figure 1), and achieve 0.450 average precision (AP) on full images, and 0.937 AP on cropped regions around dispensers.

We additionally introduce a novel, viewpoint-invariant method for 3D human pose estimation that can localize hands in the video frame to associate the actions with specific persons. Our model is a recurrent convolutional neural network that takes an image as input and outputs human pose, i.e. locations of a set of body parts. While previous methods4 have been able to accurately estimate pose from frontal views, our model uses recurrent error feedback5 in a viewpoint invariant framework to achieve strong pose estimation from arbitrary and challenging viewpoints such as a top-down view over a hand hygiene dispenser. Our method achieves 73% localization accuracy of left hands and 65% of right hands, where a detection is correct if the localization error is less than 10cm.

Conclusion

We have demonstrated the use of image-based and human pose-based computer vision methods for monitoring hand hygiene compliance. Future work involves demonstrating these methods at large scale in hospital environments, implementing real-time interventions based on the monitoring, and measuring their effect on compliance.

References

Annotation of Tumor Reference Resolution and Tumor Characteristics for Cancer Liver Stage Prediction

Wen-wai Yim¹, Tyler Denman², Sharon W. Kwan, MD²,³, Meliha Yetisgen, PhD¹,⁴
¹Biomedical and Health Informatics, ²School of Medicine, ³Department of Radiology, ⁴Department of Linguistics, University of Washington, Seattle, WA

Introduction
Despite availability of new treatments, deaths due to liver cancer, one of the leading causes of cancer-related deaths worldwide, have continued to increase in the United States. More than 80% of patients do not survive past 5 years. Various co-morbidities, lifestyle, and genetic diversity are associated with the disease; thus, optimization of treatment strategies to personal characteristics remains challenging. We target evidence-based research in this area, by concentrating on automatic cancer stage classification. The premise is that cancer stages, calculated based on individual characteristics and linked to successful treatments and outcomes, may inform future clinical decisions. Tumor information, such as number and size, are typical cancer stage parameters, which require reasoning over multiple sentences. Therefore, reference resolution, which determines which mentions in text refer to the same real world entities, is an important part in cancer stage determination. In this abstract, we present our annotation of tumor reference resolution and tumor characteristics used to calculate liver cancer stages.

Annotation
Our goal of automatic stage classification for three liver cancer staging systems (AJCC—the American Joint Committee on Cancer, BCLC—the Barcelona Clinic Liver Cancer, and CLIP—the Cancer of the Liver Italian Program) necessitates identification of a few general tumor-related stage parameters. Based on our staging systems, we identified three tumor-related relevant parameters: (1) tumor number, (2) tumor size, and (3) invasion of >50% of the liver. However, these values are not only non-explicitly stated in reports, they require reasoning over multiple sentences. For example, as shown in Figure 1, arriving at the relevant parameters (1-3) requires understanding co-referring statements and correctly aggregating.

We annotated reference resolution and tumor characteristics for a corpus of 101 abdomen radiology reports from University of Washington hepatocellular carcinoma patients, which was previously annotated for tumor events [1]. Reference resolution annotations were marked using brat, a web-based annotation software. The two types of reference resolution annotations included: (a) coreference relations, directionless link amongst two or more references of the same entity, e.g. line-29-Lesion-1 and line-38-lesion, and (b) particularization relations, a directed relation between a generalized reference to more specific reference, e.g. line-37-focal-lesions to line-38-lesion. The summarized tumor characteristics (1-3) mentioned above were annotated per document. Tumor numbers for benign, indeterminate, and unknown malignancy were also marked (instead of only malignant tumors). Thus, for Figure 1, the document annotations include the following: (#-malignant-tumors:2), (#-indeterminate-tumors:1), (#-benign-tumors:0), (#-unk-tumors:0), (size: 3.6cm), and (>50%:no).

Inter-annotator agreement for 20 documents, annotated by a medical student and biomedical student for reference resolution, showed an agreement of 0.94 F1 (MUC) for coreference and 0.84 F1 for particularization relations. Agreement for tumor characteristics was at 0.91 F1.

Future Work
We are building an automatic system for reference resolution and tumor characteristic annotation. We will include these pieces into our overall cancer stage prediction system, though our methods are generalizable to other systems.

Acknowledgements
National Institutes of Health, National Center for Advancing Translational Sciences (KL2 TR000421) and the University of Washington Institute of Translational Health Sciences (UL1TR000423).

References

Figure 1. Radiology report example
29: Lesion 1: segment 3, 3.6 x 2.3 cm hypervascular with washout
……..
35: Impression:
36: Agree with outside reports:
37: 3 focal lesions.
38: Segment 3 lesion is consistent with HCC.
39: Segment 4A/B lesion is indeterminate.
40: Segment 7 lesion is suspicious for HCC.

1650
Visualization of Topics from Twitter and Focus Groups as the Foundation for Insights about Dementia Caregiving

Sunmoo Yoon, PhD1, Niurka Suero-Tejeda1, Blake Hunter, PhD, Suzanne Bakken, PhD1
1 Columbia Univ. Medical Center, NY, 2Dept. Math Sci, Claremont McKenna College, CA

Abstract
The study aims to illustrate application of mining and visualization techniques for insights about dementia caregiving.

Introduction
Approximately 47.5 million suffer from dementia. Informal caregivers such as their family members experience burden and need support for their self-management tasks and skills including emotional, medical and role management of the individual with dementia as well as themselves. Mining social media may provide novel insights of support strategies for dementia caregivers. The study purpose is to illustrate application of mining and visualization techniques as a foundation for insights about dementia caregiving.

Methods
Using NCapture, we randomly extracted Tweets mentioning dementia (n=73564) or demencia (n=2501) from Tweet corpora collected daily from Sep 2015 to Jan 2016. We applied Latent Dirichlet allocation (LDA)-based topic modeling1 to identify topics from the Tweet corpus after preprocessing with R. The authors with a native speaker analyzed focus group data by human coding of English and Spanish verbatim transcripts and visualized the topics using network diagrams.

Results
Figure 1 displays topics from English and Spanish Tweets and focus group data.

Figure 1. Topics from English and Spanish Tweets and focus group data

Conclusion
Application of mining and visualization techniques provide the foundation for research team insights regarding dementia caregiving and self-management needs for the person with dementia and the dementia caregiver. The findings will be applied for future targeted intervention development for Hispanic dementia caregivers.

Acknowledgment: R01NR014430-03S2, NSF Institute for Pure and Applied Mathematics (IPAM)

References
Using Topic Models to Analyze Concepts for New MeSH Terms

Zhiguo Yu, MS1, Todd R. Johnson, PhD1
1The University of Texas School of Biomedical Informatics at Houston, Houston, TX

Abstract

In this poster, we present a way to use Topic Models to analyze the evolution of concepts for new MeSH terms. By computing the correspondences between the MeSH terms and the topics generated from PubMed citations published in each year, we show that there are topics that are more and more similar to the MeSH terms along the years. When the MeSH term was introduced, a correspondence score around 0.75 was reached.

Introduction

Medical Subject Heading (MeSH) is one of the source vocabularies in UMLS with the primary purpose of supporting indexing, cataloging, and retrieval of biomedical articles in MEDLINE. A small group of indexing staff at the NLM are responsible for revising and updating the MeSH vocabulary. However, NLM does not retrospectively re-index MEDLINE citations with new MeSH terms. Therefore, the timing to introduce a new concept to MeSH vocabulary is crucial. Unlike MeSH terms, which are manually created, topic models can automatically extract semantic themes based on a subset of documents. Such themes may uncover a specific set of topics for a particular domain or sub-domain, which may not exist in MeSH. In our previous work, we introduced a correspondence matrix (figure 1) between topics generated using Latent Dirichlet Allocation (LDA) and MeSH terms. Each topic generated by LDA is a distribution over all of the unique words in the corpus. We represented each MeSH term as a distribution of the words contained in the documents to which it had been assigned. A similarity score was then computed between each pair of topic and MeSH. Figure 1 provides an illustrative correspondence matrix. Here, Topic 2 is the corresponding topic for MeSH 2.

Figure 1. Correspondence Matrix.

To analyze the evolution of concepts for new MeSH terms over time, we used 3368 titles and abstracts returned from the PubMed query, ‘J Biomed Inform’[Journal] OR ‘J Am Med Inform Assoc’[Journal]’. We picked two MeSH terms, ‘Data Mining’ and ‘Electronic Health Records’, which were both introduced in 2010. We created each of these two MeSH terms’ word vectors using all the citations indexed with this MeSH term within the query results. For each year, we used LDA to generate topics based on the citations published in and before this year. Then we computed the correspondences between topics and these two MeSH terms for each year. We chose the highest corresponding (similarity) score for each year to determine whether we could find similar concepts prior to the introduction of the new terms. Figures 2 and 3 show how the highest similarity score varies and improves over time. There are topics that are more and more similar to the two terms over time. From these two examples, we see there are topics in 2008 and 2009 that have the same similarity score as the topics of 2010, in which those two MeSH terms were introduced. In both cases, a correspondence score around 0.75 was reached at the time the new terms were introduced.

Figure 2. Topic evolution for ‘Data Mining’. Figure 3. Topic evolution for ‘EHR’

References

A Pilot Evaluation of the NIH Common Data Elements for Standardizing the Data Collected in Clinical Research Studies

Marianne Zachariah1,2, Amanda L. Do, MPH1,2, Jennifer Imaa1,2, Omolola Ogunyemi, PhD1,3, Liz Y. Chen, MBA1,4, Spencer SooHoo, PhD1,5, Kevin Dawson, MD, MBA, MS1,4, Robert A. Jenders, MD, MS1,5, and Douglas S. Bell, MD, PhD1,2

1UCLA Clinical & Translational Science Institute, Los Angeles, CA; 2UCLA David Geffen School of Medicine, Los Angeles, CA; 3Charles R. Drew University of Medicine and Science, Los Angeles, CA; 4Los Angeles Biomedical Institute at Harbor UCLA Medical Center, Torrance, CA, 5Cedars-Sinai Medical Center, Los Angeles, CA

Introduction

The NIH Common Data Elements (CDEs) is an initiative spanning the NIH institutes to develop standardized elements for inclusion in research data collection instruments. Its ultimate aim is to make data collected by different research projects more comparable, so that data from multiple studies can be pooled more easily for reanalysis. One of the oldest and best-developed of the CDE initiatives, the National Institute of Neurological Disorders and Stroke (NINDS) CDE currently contains 16,000+ unique variables derived from more than 550 instruments.

For this study, we evaluated the extent to which the variables collected by a major stroke research study using REDCap, a software system used for implementing standardized research data collection, were or could be represented using the NINDS CDEs. We explored the extent to which the CDEs encompassed the data needed for the stroke study and the extent to which investigators might need help in finding appropriate CDEs.

Methods

The REDCap forms from a UCLA neurology study that utilized NINDS CDEs were examined. The study contained 1,331 REDCap fields. Of these, 468 fields were excluded from our analysis because they were duplicates of other fields that were translated into Spanish, and an additional 507 fields were excluded because they did not capture discrete subject variables (e.g. instructions for the interviewer, flow control, modifiers of other variables, etc.). The remaining 361 discrete patient variables were evaluated by four reviewers to determine whether an NINDS CDE was used or could have been used. NINDS CDEs were searched using an Excel file downloaded from the CDE website in Summer, 2015. When possible, each study variable was matched to an NINDS CDE, and the closeness of the match was categorized. To define the match process and match types, the first 129 of the 361 discrete variables were reviewed by each reviewer, in sets of 20-40 items at a time, with comparisons of match and consensus-setting meetings after each set. After consensus match definitions were agreed upon, the remaining 232 variables were then each evaluated independently by two of the four reviewers. A fifth reviewer adjudicated disagreements between reviewers on the CDE match and match type. Cohen’s κ was calculated to measure inter-rater agreement.

Results

Of the 361 discrete patient variables captured by the study, a closely related CDE was found for 117 (32%). Of these, 20 (17%) used a CDE item unmodified, 59 (50%) used a CDE with minor modifications, and 38 (32%) were items that had an unused potential CDE equivalent. Of the 244 without a closely related CDE, 56 (23%) could be matched to a partially related and 58 (24%) to a distantly related CDE. Focusing on the 232 study variables that were independently evaluated, reviewers agreed on 43% of match types, and inter-rater reliability was fair (Cohen’s κ = 0.264; 95% CI [0.201, 0.328]). In choosing a specific CDE (or no CDE), reviewers agreed for 53% of the study variables.

Discussion & Conclusion

Preliminary analysis indicates that investigators succeeded in using NINDS CDEs but that the CDEs need expansion given that a closely related CDE was found for only 32% of study variables. Although the 16,000+ CDEs are categorized into nine domains, the domains are not defined. Clearly defined domains and improved organization of the CDEs would facilitate the identification of the appropriate standard when it is available. Poor agreement between reviewers is a reflection of the cumbersome task of searching the 16,000+ variables in the NINDS CDEs that are grouped into nine undefined domains.
Reading Leadership: Improving Participation in an Introduction to Health Informatics Course

Iris Zachary, PhD, MS, CTR\textsuperscript{1}, Suzanne A. Boren, MHA, PhD\textsuperscript{1}
\textsuperscript{1}Department of Health Management and Informatics,
University of Missouri, Columbia, MO

Abstract

We evaluated the opportunity to improve participation in an introduction to health informatics course by incorporating four engagement strategies and offering in class and online discussion opportunities. Offering multiple methods of participation led to 100\% of students participating in the discussions.

Introduction

The purpose is to evaluate the process of student led discussion on peer reviewed health informatics articles by offering more than one method of participation in an Introduction to Health Informatics course. Previous students had provided general feedback that there was not enough time in class to fully discuss each article. Some students had cited barriers to successfully joining the in class discussion with their peers. We have incorporated four engagement strategies, 1) intellectual engagement by having the student pose two questions that are discussed by the class, 2) behavioral engagement by having student led discussion, 3) social engagement by engaging discussion among the students, and 4) cultural engagement by giving students different modalities they can voice their opinions either in class or online, however they feel more safe to participate and confident in terms of language barriers.

Methods

Each student in the Introduction to Health Informatics course was required to complete the reading leadership assignment. Reading leadership consisted of each student leading the discussion of an article from the peer reviewed published literature related to course topics. Students were provided with a pre-approved article list and were required to read all articles. Students wrote a 150-200 word summary and two discussion questions posted to a discussion forum in the course management system, made a three minute oral summary, and led a discussion lasting up to a maximum of 10 minutes. (Students could participate by responding to the discussion questions orally during the class discussion and by responding online in the discussion forum for up to 24 hours following class. Participation in the oral discussion in class was tracked by one of the course instructors as it occurred. Online discussion was read and tracked the next day. Comparisons were made between in class and online discussions participation rates.

Results

Thirty-nine (39) students were enrolled in the course and participated in reading leadership discussion throughout the semester. Thirty-two (32) or 82\% of the in-class reading leadership discussions lasted the full 10 minutes. All students participated in at least one of the discussion modalities: 32 students participated in discussion in class, 34 students participated in discussion on line, 27 students participated in both in class and online discussion. A total of 365 comments (average of 9.4 comments per student) were made in class and 306 comments (average of 7.8 comments per student) were made online.

Conclusion

Providing the opportunity for students to participate in the discussions orally in class and in writing online immediately following class expanded participation by seven students (18\%) to achieve 100\% of students participating in class discussion. Offering multiple methods of participation enabled all students to participate and many students to participate more than they would have with only one method.
Answering patients’ questions using expert-vetted online resources: A case study of diabetes

Yuqun Zeng¹², Liwei Wang¹, Yanshan Wang¹, Dingcheng Li¹, Xusheng Liu², Hongfang Liu¹

¹Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA, ²the Second Clinical College, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, P.R. China

Abstract
It is not clear if expert-vetted online resources cover the information needs of patients. Collected patients’ questions of diabetes, we evaluated three expert-vetted online resources (WebMD, MedlinePlus, and UpToDate) with respect to the content coverage and time required to find information when ignoring the language use difference. Among three online resources, WebMD received the highest coverage with the least amount of time in finding answers.

Introduction
An increasing number of patients seek medical information online. Expert-vetted knowledge online resources such as UpToDate have been created by medical authorities to provide new and accurate health and disease information which have a great potential to provide the right information for patients. However, it is not clear if they contain the information needed by the patients. In this study, we evaluated three expert-vetted online resources (WebMD-WMD, MedlinePlus-MLP, and UpToDate-UTD) with respect to the content coverage and time required to find information when ignoring the language use difference.

Methods
190 questions were obtained from a healthcare data analytics challenge in ICHI2015-IEEE International Conference on Healthcare Informatics. Two experts with medical background read the questions, identified keywords to be used for searching the content to answer these questions from the three resources. For each question, the time used to retrieve was recorded, and the answer on each question was assessed on a six-category rating scale ranging from 0 (none), to 5 (complete and comprehensive answer matching with question). One expert searched the three resources for all 190 questions and the results were analyzed. While the other expert searched 50 questions randomly selected from those 190 questions and the results were used to evaluate the consistency by calculating the weighted κ value (Cohen’s kappa coefficient) of scores and time in the same 50 questions from the two experts.

Results
The comparison of scores between the three groups (Figure 1) showed that three group scores were all more than 4. The score of MLP (4.02±1.04) was remarkably lower (p<0.05) than that of UTD (4.53±0.77) and WMD group (4.58±0.67). No statistical difference was observed between UTD and WMD group. In comparison with WMD (1.00±0.75), UTD (1.42±1.00) group took remarkably longer time (p<0.05) while MLP (1.07±0.74) was not statistically significant comparable to WMD (Figure 2). The kappa values greater than zero (κ>0) indicated that the two experts had the consistency for the results. Among three online resources, WebMD received the highest coverage with the least amount of time in finding answers.

References
Data-Driven System for Perioperative Acuity Prediction

Linda Zhang, BS\textsuperscript{1}, Daniel Fabbri PhD\textsuperscript{1}, Jonathan P. Wanderer MD, MPhil\textsuperscript{1,2}
\textsuperscript{1}Departments of Biomedical Informatics\textsuperscript{1}, and Anesthesiology\textsuperscript{2} Vanderbilt University, Nashville, Tennessee

The American Society of Anesthesiologist’s’ (ASA) Physical Status classification is a subjective assessment of a patient’s overall health. It is a widely used grading system for preoperative health in surgical patients, and consists of six classes of increasing severity [1]. Associations of ASA score with postoperative outcomes have been reported in literature [1]. Even though ASA score has been shown to be a predictor of postoperative outcome, there is considerable variation in the ASA classification allocation [1]. It is up to the anesthesiologists to subjectively rate the patients’ preoperative health. Though there are definitions for the class allocations in the ASA scale, the definitions do not detail the diseases and conditions for patients in each class. Despite that, the ASA scale is considered reliable. A study recent study performed in 2014 shows that the ASA scale has moderate inter-rater reliability in clinical practice [2]. The biggest discrepancy in scores is between ASA class 2 and 3.

Since ASA score is an important indicator for outcomes, it is important that scores for patients are consistent. There would be less variability if the method of evaluating ASA score was not so subjective. In this study, we attempt to solve the problem of predicting the ASA score of patients using a reproducible, data-driven automated system. Using data from the Vanderbilt Perioperative Data Warehouse, which houses data from the perioperative environment at Vanderbilt University Medical Center, we develop a model to predict the ASA score of a patient. We use supervised machine learning techniques to develop a system that mines the underlying factors that determine if a patient belongs in ASA class 1-2, or ASA class 3-5.

The machine learning techniques that we test in our study were logistic regression, random forest, and deep neural networks. We first chose features for the ASA prediction classifier. The features were age, body mass index (BMI), service category that the patient was in, if the patient previously had surgery, home medications, prior inpatient ICD9 codes and history of outpatient ICD9 codes. We incorporated temporality into the ICD9 features by sectioning codes into stretches of time for which they were placed. From these features, we tested different combinations of features and created a combined classifier. The final features we selected were age, BMI, service, medication class, inpatient ICD9 hierarchy, and outpatient ICD9 classes. We also compared the combined classifier to classifiers using a single feature (i.e. ASA prediction with only age). We used logistic regression and random forest to test these, measuring performance with 5-folds cross validation (Figure 1).

![Figure 1. Logistic regression and random forest results for single categories of features and combined categories](image)

We then used deep neural networks to make predictions from the features in the combined classifier. The best network that we developed used 3 layers and 800 units, and achieved an AUC of \(0.874\pm0.0024\). In conclusion, we have found that it is possible to predict the AUC of patients using data from the EMR and achieve a high AUC.

References

An Integration of Cooperative Learning into Undergraduate Medical Informatics Class with Service Learning Components

Jiangyuan Zhou, PhD1, Duo (Helen) Wei, PhD2

1 School of General Studies, Stockton University, Galloway, NJ 08205, USA; 2 Computer Science and Information Systems – BUSN, Stockton University, Galloway, NJ 08205, USA

Abstract
Service learning module in undergraduate Medical Informatics class provides educational and outreach activities and builds a mutually beneficial relation with healthcare community partner. However, student learning outcome result shows that students were lack of teamwork skills with their classmates when participating community service. In response to this problem, a cooperative learning method was introduced into the Medical Informatics class with service learning components. Students are encouraged to work together as a team to improve community participants’ awareness of the usage of Smartphone Apps to assist in managing their chronic diseases. The pre- and post- class surveys show positive results. Students welcome this method and confirmed the benefit of this cooperative learning approach in Medical Informatics education.

Introduction
Service learning module was introduced to an undergraduate Medical Informatics class in a four-year state university to develop community awareness, communication, and critical thinking skills in 2012. An assessment tool was developed with 20 items on Engagement, Learning, Communication, Collaboration and Critical Thinking to assess the students’ performance in this service learning course by collecting the observations from community partners, instructor, classmates and student themselves. The results show that the main barrier for this medical informatics class was the teamwork skills among classmates. So a cooperative learning method was introduced into the Medical Informatics with service learning component in 2015. The service learning project was designed to have a common goal for the entire class to increase community awareness of using Smartphone Apps to assist in managing chronic disease. Students are assigned to small groups with various tasks, such as organizing workshop, developing survey, creating posters, and collecting/analyzing data. They worked together as a team and wrote self-reflection papers. At the end of the semester, students are expected to present their tasks and their self-reflections. The study aims to evaluate the effects of this Medical Informatics course with the introduction of cooperative learning[1] method.

Methods
Student performance scores on teamwork skills of four classes from 2012 to 2015 (in total of 114 students) were compared. Variance analysis was used to determine the difference among four groups. Pre- and Post- course surveys on the integration of cooperative learning method into the Medical Informatics class were conducted to determine students’ perceptions about the integration of cooperative learning with medical informatics class with service learning components.

Results
The average scores of the one groups (2015, 35 students) after the integration are higher than the other three groups (2012, 2013, 2014, in total 79 students). Variance analysis showed there were difference among the four groups (p<0.05). Pre- and Post- survey shows that “integrating cooperative learning method in medical informatics class is helpful.” With students’ average rating of 9.4 out of 10.0.

Discussions
Medical informatics is interdisciplinary in nature. Therefore, it make sense to seek a way to let students work together to achieve a common goal. The integration of cooperative learning into Medical Informatics with service learning components seems to be a useful approach to foster teamwork spirits. A large sample with controlled trial is desirable to confirm the effects. Also, the integration of cooperative learning into a Medical Informatics without service learning components is also worth trying.

Extracting Laboratory Eligibility Criteria Data Elements from IRB Protocols Using Natural Language Processing

Vivienne J. Zhu, MD, MS, Elizabeth A. Marshall, MD, MS, Randall W. Alexander, Sherly Roy Yesudhas, Leslie A. Lenert, MD, MS
Biomedical Informatics Center at Medical University of South Carolina, Charleston, SC

Abstract: We developed and evaluated the performance of a natural language processing (NLP) approach for translating laboratory eligibility criteria within IRB protocols into a structured format. A total of 718 IRB approved clinical trial protocols were used to iteratively develop lexicon and NLP algorithms to extract data elements (lab test name, equation, value, and unit). By manual review of 180 testing documents, our NLP approach scored 94.5% for precision, 90.2% for recall, and 92.3% for F-measure.

Background: Laboratory eligibility criteria are important benchmarks for identifying potential patients to participate in clinical research. However, laboratory eligibility criteria are commonly documented as narratives within research protocols. Usually, manual review of these protocols is required to translate such free text criteria into computerized programs in order to query medical records to identify potential subjects. A potential approach to automate the process of screening for eligibility is to use natural language processing (NLP) techniques to extract relevant data elements from free text laboratory eligibility requirements for comparison to existing structured data. In this study, we evaluate the feasibility and performance of NLP in this context.

Method: The data resource for this study is research eligibility criteria records from the electronic Institutional Review Board (eIRB) system at the Medical University of South Carolina (MUSC). A total of 898 research criteria (inclusion and exclusion) records from 898 complete or approved MUSC IRB protocols for clinical trials between 2010 and 2015 were extracted and prepared for the NLP analysis. We used commercial NLP software (Linguamatics I2E version 4.3, Cambridge, United Kingdom) licensed by MUSC to index, parse, and query each research criterion. We developed a lexicon and a set of NLP queries to capture semantic and syntactic representations of laboratory requirements from free text IRB eligibility criteria for four data elements: 1) protocol number; 2) equation for both mathematic symbol (e.g. “<,” “>,” “=,” etc.) and text presentation (e.g. “greater than,” “less than or equal to,” “at least,” “exceeds,” etc.); 3) lab values accounting for both numeric values and text representations (e.g. “10,000,” “3X,” “two times,” etc.); and 4) lab result unit (e.g. “mg/dl”, “upper limit of normal range,” “ULN,” etc.). A random sample comprised of 75% of the studies’ research criteria (718 documents) was used as training data to iteratively develop the NLP lexicon and queries. The remaining 25% of records (180 documents) were manually analyzed as a gold standard to evaluate the NLP strategy’s performance. We report standard measures: precision (=true positive/(true positive + false positive)); recall (=true positive/(true positive + false negative)); F-measure (=2* (precision * recall)/(precision + recall)).

Results: The top three specified therapeutic areas for MUSC IRB approved clinical trials are cancer (278), cardiovascular (93), and mental health (58). The NLP algorithm processed 180 testing documents within 0.2 seconds. It identified 293 laboratory eligibility criteria from 93 testing documents, and 87 documents had no information about laboratory eligibility criteria. Manual review of all four data elements for each laboratory criterion confirmed 277 positives, and identified 30 criteria that the NLP algorithms did not capture. The overall performance was 94.5% for precision, 90.2% for recall, and 92.3% for F-measure. Table 1 displays the NLP performance of the most common laboratory eligibility criteria.

Table 1. NLP performance for the most common laboratory eligibility criteria

<table>
<thead>
<tr>
<th></th>
<th>Bilirubin</th>
<th>Platelet Count</th>
<th>Creatinine</th>
<th>Creatinine Clearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of NLP identified positives</td>
<td>50</td>
<td>47</td>
<td>41</td>
<td>18</td>
</tr>
<tr>
<td>Precision</td>
<td>98.0%</td>
<td>89.4%</td>
<td>90.2%</td>
<td>94.4%</td>
</tr>
<tr>
<td>Recall</td>
<td>90.7%</td>
<td>87.5%</td>
<td>92.5%</td>
<td>89.5%</td>
</tr>
<tr>
<td>F-measure</td>
<td>94.2%</td>
<td>88.4%</td>
<td>91.4%</td>
<td>91.9%</td>
</tr>
</tbody>
</table>

Conclusion: This study demonstrates that NLP can robustly and accurately extract relevant data elements of laboratory eligibility criteria from free text IRB protocols into a structured format for subsequent comparison to coded data in an electronic health record system for automated eligibility screening.
Evaluating Acceptability and Efficacy of Antidepressant Medications using Patients Comments in Social Media

Maryam Zolnoori1, MS, Timothy B. Patrick2, PhD, Mike Conway3, PhD, Anthony Faiola4, PhD, Jake Luo5, PhD,
1,2,5 University of Wisconsin-Milwaukee, Milwaukee, WI
3University of Utah, Salt Lake City, UT
4Indiana University, Indianapolis, IN

Problem Statement
Patients with depression use the online messaging boards to post self-reports of side effects, effectiveness, and in general their satisfactions for antidepressants. The self-reports of patients experiences provide information that might not be captured in clinical trials. Because of the potential importance of such self-reports, the International Society of Drug Bulletins asserted in 2005 that “patient reporting systems should periodically sample the scattered drug experiences patients reported on the internet”. (Leaman et al., 2010).

Purpose of Study
The purpose of this preliminary study is to evaluate the efficacy and acceptability of the new-generation antidepressants for treatment of depression using patient self-reports on online messaging boards. The study is significant because depression is a periodic, incapacitating condition that is one of the major reasons for disability in developed countries. Approximately 6% of adults annually are affected by depression that is a major cause of suicide and about 9 percent of Americans show symptoms of desperate, regret, and sadness that result in depression (CDC 2014). A meta-analysis was conducted by Cipriani et al. (2009) to rate 11 new-generation antidepressants based on efficacy (treatment the signs and symptoms of depression) and acceptability (the likelihood of continuing an antidepressant medication by a patient). According to the study, the 11 antidepressants ranked from the highest efficacy to the lowest are Remeron, Lexapro, Effexor, Zoloft, Celexa, Wellbutrin, Paxil, Savella, Prozac, Cymbalta, and Luvox. Based on the measure of acceptability, the 11 antidepressants that ranked from the highest to lowest acceptability are Zoloft, Lexapro, Wellbutrin, Celexa, Prozac, Savella, Remeron, Effexor, Paxil, Cymbalta, and Luvox. Since clinical trials, by their nature and purpose, are focused on a limited number of participants selected by inclusion/exclusion criteria (such as demographics, medical condition, and diagnosis), this study may not provide adequate indications of the efficacy and acceptability of antidepressants.

Methods
The methodology of this research consists of two main phases. In the first phase, we used technique of content analysis to measure efficacy and acceptability of the antidepressant medications, and identify side-effects of the 11 new-generation antidepressants. In the second phase, we are going to apply techniques of association rules and co-occurrence to automate the process of extracting the side-effects and measuring the acceptability and efficacy of the antidepressant medications. The output of the first phase of this research is demonstrated in the results section.

Results
The preliminary result of the study showed that, in contrast with the results of the research conducted by Cipriani et al. (2009), the 11 antidepressants ranked, according to patient comments, from the highest efficacy to the lowest are Lexapro, Celexa, Paxil, Luvox, Effexor, Prozac, Zoloft, Wellbutrin, Cymbalta, Remeron, and Savella. In addition, based on the acceptability measure, the 11 antidepressants ranked from the highest to the lowest are Celexa, Lexapro, Prozac, Luvox, Wellbutrin, Zoloft, Paxil, Effexor, Cymbalta, Savella, Remeron. The result indicates, for example, that though Remeron has the highest efficacy in the meta analysis of clinical trials, it is ranked as 10th least efficacious according to the patients comments. In addition, while Zoloft has the highest acceptability according to the meta analysis, the patient comments showed that Celexa has the highest acceptability, not Zoloft.

Conclusion
Analyzing patient comments about antidepressant medications shows that there is important difference between commonly prescribed antidepressants for both measures of efficacy and acceptability. Based on the patients comments Lexapro and Celexa might be the best choice when starting treatment for depression in adults because they have the highest rating for efficacy and acceptability.

References
Personalized Heart Disease Risk Manager: A Tool for Patients and Clinicians to Manage Cardiovascular Risk

Raja A. Cholan, Jennifer A. Pacheco, Gene Ren, Laura Hickerson, MS, MLS
Dept. of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR

Acknowledgements: Aditya Rane, Experience Designer

Summary
Heart disease is the leading cause of death in the United States with over 600,000 people dying every year. To help with prevention, the American Heart Association (AHA) and American College of Cardiology (ACC) have developed tools which include risk algorithms and guidelines for healthcare providers to assess atherosclerotic cardiovascular disease (ASCVD) risk. Several of these tools provide treatment recommendations for aspirin, statins, and blood pressure medications, and calculate a patient’s 10 year risk of having heart disease or stroke. In this proposal, we describe the ‘Personalized Heart Disease Risk Manager’ (PHDRM) which aims to deliver personalized healthcare for managing ASCVD risk by (1) incorporating a patient’s environmental, lifestyle, and family history risk factors, (2) highlighting relevant genetic tests for family history associated with ASCVD risk from ClinGen and the Genetic Testing Registry, (3) adding relevant pharmacogenomic tests for aspirin, statins, blood pressure medications, and smoking cessation medications from the Clinical Pharmacogenetics Implementation Consortium (CPIC); and, (4) enhancing the patient-centeredness of education materials in existing ASCVD risk calculators. The PHDRM would act as a decision aid and help clinicians engage patients in targeted conversations regarding treatment options, genetics, behaviors, and environmental factors. It would also help clinicians stratify which patients should get genetic testing for ASCVD risk and/or pharmacogenomics. The PHDRM will help patients understand their cardiovascular risk factors and enhance their ability to make informed decisions. This solution can be incorporated into the point-of-care through bi-directional SMART on FHIR integration via electronic health records and patient portals. An interactive prototype of the PHDRM is available at https://invis.io/957TONK6Q

Background

Problem Definition: Importantly, several ASCVD risk factors are modifiable through lifestyle change or drug therapy, including smoking, high blood pressure, dyslipidemia, diet, exercise, and alcohol consumption. Therefore, patient engagement to achieve successful preventative measures is essential. For ASCVD risk factors not modifiable, such as race/ethnicity, family history and genetics, awareness of high-risk genetic factors is still important for prevention. Several existing ASCVD risk models and associated tools enable identification of select patient ASCVD risk factors, however few consider family history as part of the risk assessment. Additionally, ASCVD risk tools commonly only address risk identification but lack functionality to engage patients in shared-decision making, promote healthy lifestyle over the long-term, or provide resources to that end. Many ASCVD-related guidelines, including those from the American Heart Association (AHA) and American College of Cardiology (ACC), share in common a healthy lifestyle as a foundation for ASCVD prevention. However, functionality to address this healthy lifestyle foundation is lacking in many ASCVD risk tools. In their recent review of the scientific literature relevant to the 2013 ACC/AHA guidelines, Cainzos-Achirica and colleagues propose that healthy lifestyle, and the tools to achieve it, should be “thoroughly addressed in a clinician-patient discussion.” An interactive ASCVD risk assessment tool will not only help identify patient ASCVD risk factors, but also facilitate the clinician-patient discussion about healthy lifestyle, encourage shared-decision making, enable patient-goal setting and progress monitoring, and provide patient resources to help adopt healthy habits. This aligns with the tenants of the National Institutes of Health’s ‘Precision Medicine Initiative’ which places an emphasis on tailoring prevention and treatment efforts that take individual variability into account.
The AMIA Challenge: Our team therefore focused on designing a web-based cardiovascular risk calculator and patient decision aid tool to facilitate conversations between patients and clinicians. Using information gathered from 14 primary care physicians (PCPs) and cardiologists at Oregon Health & Science University (OHSU), Northwestern University, and Mayo Clinic, and 10 potential patients, as well as consultation from experts in shared decision making, data visualization, genetic counseling, and experience design, we propose the ‘Personalized Heart Disease Risk Manager’ (PHDRM). This tool focuses on engaging patients and clinicians in targeted conversations thereby informing the decision-making of both clinicians and patients through enhanced understanding of genetic profiles, treatment, environment, and lifestyles related to improved cardiovascular risk. The PHDRM will have a clinician-only login where clinicians can see clinical, evidence-based information related to genetic information, with the aim of improving clinical interpretability.

Solution Design and Development

Design Process: The Agency for Healthcare Research and Quality (AHRQ) is funding the EvidenceNOW grant initiative that is focused on helping thousands of primary care practices across seven regional cooperatives improve patients’ cardiovascular health, with the aspirational goal to meet national targets on four cardiovascular clinical quality measures. One of our team members is a Health IT practice facilitator working with practices in Oregon on data extraction and quality improvement interventions for these cardiovascular outcomes. The inspiration for the designing the PHDRM came through a qualitative needs assessment (phase 1) through his interactions with clinicians. We identified the following common themes during our phase 1 needs assessment: (1) clinicians and patients do not always agree with blanket guidelines; there needs to be room for clinical judgement and patients’ values and preferences (e.g., patients whose blood pressure readings are high during office visits, but within range in home readings); (2) some clinicians prefer to focus on lifestyle modifications first, such as dieting and exercise, and if there is still a need for improvement after these interventions, then consider treatment options; (3) current risk tools embedded within EHRs do not record ASCVD risk percent as structured data, limiting the ability to write queries and generate reports for stratification; and (4) accessing patient decision aids embedded in EHRs takes a long time, but the subsequent conversations with “defiant” patients is useful.

Once our initial framework was developed, we conducted a more formal needs assessment (phase 2) with additional provider types, and researchers at multiple institutions, and also with potential patients, in which we asked for feedback on our initial concepts and lo-fidelity wireframes (see appendix A for the questions we asked). We met with representatives of the Knight Cardiovascular Institute, Department of Medical Informatics and Clinical Epidemiology, the Oregon Rural Practice-based Research Network at OHSU; representatives from Cardiology, Internal Medicine, Genetic Counseling at Northwestern University; and a cardiologist from Mayo Clinic. From these providers, we learned that while the interviewed providers used the existing ACC/AHA estimator, some were hesitant to use genetic testing, yet they did think that the existing tools could be improved in order to help patients understand their current calculated risk, optimal risk, and be more motivated to make lifestyle changes, to comply with treatment plans, and document their goals. These clinicians would like to see genetic testing used when there is not enough other information to make an informed decision, or when patients are at high risk for well-established genetic risk factors in the Genetic Testing Registry like familial hypercholesterolemia. They also wanted a tool that would allow them to use the limited time they have during appointments to focus on shared decision making for treatment and lifestyle modification plans.

The findings from each of the patient interviews reveal differences in perception of modifiable and unmodifiable ASCVD risk factors. The modifiable risk factors include sociodemographic, lifestyle, and behaviors. For example, two patients were aware that the lack of exercise and consumption of unhealthy foods are harmful to their health but did not have a plan to make changes. Most patients were interested in genetic testing, with most
of those wanting all genetic test results regardless of whether treatment is available or genetic risk is currently fully understood, although at least 1 patient was more interested in genetic testing for another disease area (dementia), not ASCVD. Most patients wanted to know more about what lifestyle modifications they could make to lower their risk and most felt they could use more education and support in order to make those changes.

Following both phases of the needs assessments, we created wireframes using software called Sketch, and created a clickable prototype in a program called Invision. Our prototype simulates the PHDRM as it might appear during a clinician-patient interaction for a fictional persona. Text about patient education was curated by reviewing several existing materials from the Office of the Patient Care Services National Center for Health Promotion and Disease Prevention, Million Hearts, National Institutes of Health, National Heart, Lung, and Blood Institute, Healthwise, and Mayo Clinic. ‘Calculated’ and ‘Optimal’ 10-year ASCVD risk calculations presented in our prototype were generated using the ACC/AHA algorithm for the purposes of our demonstration. We relied on a paper written by Safarova and Kullo which outlines how to screen and treat patients with familial hypercholesterolemia; our patient persona that we created has a family history of high cholesterol, and therefore likely has familial hypercholesterolemia.5

Solution Description: Our proposed solution, as aforementioned, is a web-based heart disease risk calculator and patient decision aid tool named, ‘Personalized Heart Disease Risk Manager’ (PHDRM). The PHDRM tool has four main components: (1) a risk calculator using the 2013 ACC/AHA guideline for the assessment of cardiovascular risk which gives users the ability to set targets and visualize calculated risk compared to optimal risk, (2) patient education materials about family history and genetic screening, treatment, lifestyle, environment for promoting heart health, (3) documentation of patient goals on various pages of the tool, and (4) information about pharmacogenomics when patients and clinicians may want to consider testing for certain genes associated with medications and familial conditions.6

We envision this solution being used during clinical care as a shared decision making tool for primary prevention of ASCVD, primarily for patients aged 40-70, or those otherwise deemed to have a potential higher risk of ASCVD (e.g., patients with diabetes and elevated LDL), as follows see Appendix B.1. for flowchart. First, a week before a patient has a regular check-up appointment with their PCP or cardiologist, the patient will receive an appointment reminder from their EHR patient portal, such as Epic’s MyChart ©, and in that message, there will be instructions to complete any questionnaires before the appointment, with a link to the PHDRM tool as one of those questionnaires (see Appendix B.2. for an example of what this looks like in Epic’s MyChart©). Then, when the patient uses the PHDRM tool, demographic and clinical data will be filled in for the patient from the EHR (using FHIR), the patient will simply fill out their family history and lifestyle behaviors, and click ‘Calculate’ to see their ASCVD risk score. The patient can then review their results, read the educational materials, draft goals to improve their ASCVD risk and write down any questions they may have for their doctor in the PHDRM tool. On the day of their medical visit, the clinician can preview the results and questions in the EHR via the ‘clinician login’ when preparing for the appointment. Additionally, some information, such as updates to family history for example, could be updated in the patient’s record in the EHR via integration through FHIR. During the patient encounter, the clinician and patient can focus on discussing the results and making shared decisions to reduce the patient’s ASCVD risk. If further testing or treatment is ordered, the PHDRM tool will be updated with those test results and treatments as they are entered into the EHR. The patient or provider can then see the updated risk score based on the new treatments and any lifestyle changes that are made. At the patient’s next encounter, the patient and provider can repeat the process again, this time reviewing if goals were met, and if not, discuss any barriers to those goals, and make new goals as needed. In addition, as the CPIC and ClinGen curate new guidelines for pharmacogenomics and other genetic testing for medications and ASCVD risk, we will update our PHDRM tool to incorporate these, with feedback from physicians and potential patients.
Alternative Solutions and Solutions Comparison

Based on our interviews, clinicians who do use ASCVD risk tools at the point of care tend to rely on the ACC/AHA Risk Estimator, or they use ASCVD risk tools embedded within their EHRs that use the ACC/AHA algorithm (e.g., the Kaiser Permanente National Guideline Cardiovascular Risk and Dyslipidemia Management Clinician Guide 2015). The ACC/AHA algorithms require age, gender, race, total cholesterol, HDL, systolic blood pressure, diastolic blood pressure, treatment for blood pressure, diagnosis of diabetes, and smoking status in order to calculate risk. Table 1 summarizes the alternative solutions along with the strengths and weaknesses for each.

Table 1. Comparison of ASCVD risk tools for primary prevention

<table>
<thead>
<tr>
<th>Possible Solution</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Personalized Heart Disease Risk Manager</td>
<td>Potential integration with EHRs/patient portals via FHIR.</td>
<td>Potentially time-consuming data entry for patients.</td>
</tr>
<tr>
<td></td>
<td>Incorporates genetic testing guidelines from ClinGen/CPIC.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assimilates family history and lifestyle choices.</td>
<td>Many clinicians are not currently trained to understand genomics information and guidelines.</td>
</tr>
<tr>
<td></td>
<td>Patient-friendly education materials which are engaging, easily understandable, and actionable.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clinician view with clinical, evidence-based information.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Personalized goal setting, and self-management support</td>
<td></td>
</tr>
<tr>
<td>ACC/AHA ASCVD Risk Estimator</td>
<td>Interactive, provides calculated and optimal CVD risk, and provides lifetime CVD risk as well as 10-year CVD risk</td>
<td>Optimal risk factors are not personalizable.</td>
</tr>
<tr>
<td>Mayo Clinic - Statin Choice Decision Aid</td>
<td>Provides clinicians with text to copy/paste into the progress note indicating that a decision aid was used. Incorporates cost, daily routine, other benefits/risks for patients to consider</td>
<td>Does not incorporate family history, genetics, lifestyle.</td>
</tr>
</tbody>
</table>

Implementation and Dissemination

We have spent over four months developing our current prototype with the input and feedback from both provider and patient stakeholders. We will continue to iteratively improve our tool based on further needs assessments and feedback, and we will add additional features such as improved data visualization regarding risk, and links to other resources. Future builds would include a working standalone version first, then a version that integrates with EHRs via FHIR (both reading data into the tool and writing data back to the EHR data such as risk score and family history), and a version that reads from personal patient mobile health devices via SMART on FHIR (e.g., smartphones, smart watches, home blood pressure monitors).

To disseminate our tool to others we will work with stakeholder groups like eMERGE (electronic Medical Records and Genomics) and Precision Medicine Initiative enrollment centers, organizations such as the ACC, AHA, and AHRQ to promote and get feedback on our tool. We will also work with EHR vendors such as Epic and Cerner to integrate our tool into EHRs.

Evaluation

The proposed PHDRM tool will be evaluated through usability testing, and comparative effectiveness evaluation of similar ASCVD risk tools. Usability testing has been recommended as an important part of the evaluation of health risk communications, and is a recognized method to achieve user-centered improvements in health information, and research-based improvement in web information resources. Usability testing will follow recognized methodologies, allow us to observe actual tool users, and evaluate if the PHDRM tool meets user
needs. Results will enable identification of potential usability problems, and inform redesign and / or improvement of the PHDRM tool.

Our comparative effectiveness analysis will compare the effectiveness of the PHDRM tool against other tools, such as the ACC/AHA risk tools, Mayo Clinic Decision Aid, and potentially other decision aids from Healthwise. The outcome of interest for this comparative effectiveness analysis would be scores from the Decision Conflict Scale, to assess whether the PHDRM helps patients make decisions about their heart health.  

**Conclusion**
Cardiovascular disease is the leading cause of death in the United States and there is growing demand for preventive efforts. There is also demand for tools that help clinicians with parsing through and handling the complexity of clinical and biomedical “Big Data”. We believe that a tool like the Personalized Heart Disease Risk Manager has the potential to deliver personalized cardiovascular healthcare, to help engage patients with clinicians and the data to make informed decisions from knowledge gained from genetic, environmental, and lifestyle information.

**References**

CareMax: A Patient-Centric Cancer Treatment Decision and Calendar Tool

Jared Erwin¹, MS, William Kearns¹, Wayne Liang¹, MD, Carolyn A. Paisie¹, PhD
¹Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

Abstract

For pediatric patients with advanced cancers, deciding among treatment options can be a difficult task. Patients and caregivers rely on the advice of their providers who utilize both clinical data and the patient’s wishes and goals to generate personalized treatment recommendations. We present CareMax, an application which supports communication, shared decision-making, and personalized medicine by visualizing treatment options as calendars, allowing direct comparison among treatment options, and matching treatment options to patients’ values and goals.

Prototype: http://kearnsw.com/CareMax/

Introduction

Precision oncology frequently refers to simply matching patients to treatment options based on their cancer type and genetic profile. However, in actual clinical practice, personalized treatment decision-making is a complex, multifaceted endeavor which includes not only the patient’s medical status and tumor genetics, but also their lifestyle, social support, values, goals, and wishes. In pediatric oncology, decision-making takes on additional layers of complexity due to the special role that caregivers (e.g. parents) play in decision-making. The caregiver role also varies across diverse family structures and changes with the age and developmental stage of the patient. Choosing among cancer treatment options involves the complex task of understanding and comparing numerous features (e.g. specific medications, effectiveness, specific side effects, logistical issues), each with a different level of prioritization for different patients. Therefore, pediatric oncology is a particularly well-suited domain to demonstrate the complexities involved in implementing precision medicine.

At Seattle Children’s Hospital (SCH), patients, caregivers, and pediatric oncology medical providers frequently engage in shared decision-making in order to select a cancer treatment regimen among multiple options. These decision-making episodes may occur during a routine clinical encounter, or they may occur during a “care conference”, a meeting dedicated to discussing the implications of treatment options in order to make a decision. Decision-making is often difficult due to complex informational requirements as well as the emotional and psychosocial context of the cancer diagnosis.

Patients, caregivers, and providers have different information needs. Existing informational resources, such as treatment “roadmaps” (a paper form which guides providers in administering the regimen) and chemotherapy side effect handouts, are not in the optimal form to effectively inform decision-making. Existing resources regarding treatment regimens are generally written for medical providers, and they do not provide customized information for specific patients. Any personalized information for specific patients, such as a treatment calendar or a care conference summary document, must be generated manually by providers. Other information, such as the patient and caregivers’ goals and values, is difficult to obtain. Finally, the process of matching patient characteristics, including their clinical findings, genomic alterations, or personal values and goals, to treatment options is a time-intensive, manual process.

Our design provides an innovative approach to precision medicine in oncology by focusing on effective presentation and comparison of complex information across cancer treatment options, supporting patient/caregiver/provider engagement and communication in the shared work of treatment decision-making, and matching patients to personalized treatment options based not only on their cancer subtype, but also their values, goals, and wishes. We will pursue a multi-phase implementation approach, initially focusing on the highest priority functionality based on user needs, then progressively adding functionality with future releases. In phase one, we will implement a treatment calendar template creation tool, a treatment calendar template management system, a patient-specific treatment calendar tool, and a treatment timeline comparison tool. Future releases will implement a patient reported outcomes (PRO)-based regimen rating system, a patient goals and values capture tool, and a patient-to-regimen matching tool.
Our Solution

We initially identified shared decision making for pediatric cancer treatment options as our general domain area of interest. We then utilized semi-structured interviews with medical providers in pediatric oncology to identify user needs, set project goals, and finalize key functionality requirements.

If a patient has advanced cancer, defined as refractory disease (i.e. does not respond to initial treatment) or relapse (i.e. initially responded but then returned), it is often difficult for clinicians to determine which treatment regimens should be recommended. Options may include aggressive experimental therapy with unclear benefit, less aggressive “palliative” chemotherapy without expectation of cure, or palliative care with no chemotherapy. However, there are often no standardized treatment options, and the benefits of experimental therapies are often unknown. Expected prognosis is often low regardless of which regimen is chosen. In some cases, the side effects of cancer therapy are more severe than symptoms the patient may experience if pursuing no treatment.

From interviews with various providers, a common theme was the potential usefulness of a tool that could help patients with advanced cancers, caregivers, and providers make shared treatment decisions. Providers would find it helpful to have a tool which obtains and communicates to the provider the patient and caregivers’ values, goals, and wishes prior to the meeting. They also felt that a tool which organizes treatment options and creates customized treatment calendars would enhance communication. Finally, providers emphasized the importance of having a tool that would not add to their workload, interfere with patient treatment and care, or interfere with the sensitive nature of treatment discussions in advanced cancers.

Prior to meeting with patients and caregivers to make treatment decisions, providers do not know what decision a particular patient or caregivers may make. Values, goals, and wishes, crucial for making treatment decisions, are elicited in real time. Thus, it is difficult determine what written information to prepare in advance. The majority of providers interviewed (7 of 8) did not use any electronic tools to match patients to treatment options or to assist with information retrieval, processing, or decision-making. 1 provider reported searching ClinicalTrials.gov, a database on nationally-funded clinical trials, but they commented that the website was difficult to use. When a treatment option is selected, some providers may choose to generate a custom treatment calendar using standard desktop spreadsheet software, which is then provided to families at a later meeting. However, these calendars are created after the treatment decision and therefore do not play a role in informing decision-making. Finally, some providers may create a written care conference summary which documents what was discussed at the care conference, then provide this to the patient at a later date. Most patients do not receive a calendar or a care conference summary.

Due to privacy considerations, we were not able to interview patients for our first design iteration. Therefore, we relied on literature to provide insight on patients’ information needs. A survey of parents of children with cancer listed nine different information areas in order of priority: (a) treatments and tests, (b) cure, (c) caring for my child, (d) emotional impact, (e) side effects, (f) physical impact, (g) disease, (h) coping with painful procedures, and (i) impact on the family. The calendar functionality aims to address information around the treatment and tests (as far as frequency and duration) and help with caring for the child by giving a tool that makes adherence to the schedule easier.

We pursued an iterative design process which resulted in an interactive prototype. We then conducted focus groups with the same provider participants to demonstrate our prototype and obtain feedback. Overall, participants felt that the prototype has the potential to enhance shared decision regarding cancer treatment options by visualizing complex information and improving communication. Participants emphasized that the tool should focus on patients with advanced cancers due to high decisional complexity in those settings. Participants suggested interface modifications such as colors, layout, textual descriptors, legends, and icons. Lastly, participants suggested and prioritized additional functionality being considered for our tool. For example, participants indicated that having an interface to search treatment options at ClinicalTrials.gov and Cancer.gov was low priority. However, matching patient values to treatment characteristics would be desirable. We have incorporated their feedback into our current prototype and in our plans for future functionality.
We sought to capture patient and caregiver values, goals, and wishes in a quantitative way, which can then be used to algorithmically matched patients to treatment options. We interviewed content area experts (i.e. pediatric oncologists and pediatric palliative care specialists) and conducted a literature search in order to identify survey instruments which could potentially serve as the basis of this algorithm. We were unable to find an instrument which applied directly to our application’s goal. However, we did identify several instruments which could inform the design of our algorithm. Two instruments that are used frequently to support decision making for pediatric cancer patients in palliative and hospice settings are the Decision-Making Tool (DMT)\(^2\) and 5 Wishes\(^1\). Both tools use a qualitative, free text format to capture patient values and preferences, which does not lend itself easily to quantification for matching purposes. 5 Wishes also was designed to be an advanced directive for end of life care, which is a different situation than deciding between cancer treatments. However, the categories of variables captured in these 2 instruments were informative in our design. In the PediQUEST trial, patients with advanced cancer and their parents were administered validated, quantitative patient reported outcome (PRO) and health related quality of life (HRQoL) instruments (e.g. PQ-MSAS\(^4\) and PedsQL\(^5\)) using an electronic system\(^6,7\). Patients and parents were then provided an electronic dashboard which quantified PRO and HRQoL summative measures into 4 domains: Physical, Emotional, Social, and School. The PediQUEST study found that it was feasible to have patients (depending on age) or their parents fill out these instruments on their own in an electronic system; patients, parents, and providers reported being satisfied with receiving summative results via the electronic dashboard. We will pursue the same strategy in the design of the treatment matching functionality, and we will base our algorithm on a modified version of P-Q-MSAS and PedsQL, while incorporating elements of the DMT and 5 Wishes.

### Implementation and Evaluation Plan

There will be four main feature areas and entry points to our initial version of the system. The first area is creating a treatment calendar template which is based on paper-based treatment roadmaps currently in use at SCH. Currently, roadmaps are created through joint effort between physicians and pharmacists for best available therapy (non-research) regimens and clinical trial protocols. The roadmap includes which drugs are to be given at what doses, which tests and procedures to perform, and on which day. This paper based information will be entered into our system by a user we represent by the persona Anne (see Supplementary Material for more information on Anne). Anne directly opens the system and is authenticated using Seattle Children’s credentials over OAuth 2.0. She has access to the currently created templates which she can update, or she can create new templates. Right now Anne is representative of Nurse Practitioners at SCH, as they are the team members currently creating calendars. However, as no patient data access is required, this role could be fulfilled by someone who is not a provider which would lessen the work for providers.

The second area is creating one or more specific treatment calendars for a given patient. The persona performing this task is Ruth. She represents Nurse Practitioners, residents, and even fellows or attendings at SCH who create calendars for their patients. The entry point into the system is different for Ruth as she is doing this task in the context of a given patient. A patient’s record will be accessed in the Cerner electronic medical record (EMR) system as normal and then a new link in that system will launch our application. Patient context and user authentication will be passed to our application with OAuth 2.0 tokens. Our application will launch to calendar view with the patient banner at the top: name, birthdate, and medical record number (MRN). Any existing patient treatment calendars will be available for display and edit. A new treatment calendar for a patient can be created based on an existing template (previously created by Anne). Templates can be found by searching key features (name, therapy, clinical trial, and others) or browsed in a tree structure which is similar to the current organization of therapy roadmaps. Once a start date for the therapy is indicated, a calendar for the patient will be created with default dates based on the template. Ruth can then customize this calendar as necessary.

The third area of functionality is very similar to the second in that Ruth will open treatment calendars for a particular patient and access them via the link on Cerner as before. However, the difference is in the display and purpose of access. This area of use is to display the calendars in a meeting with patients and their families for the purpose of understanding the treatment plan and selecting between multiple treatment options. Multiple views of the treatment schedule will be available. A side by side view of typical calendar displays will allow comparison as well as a horizontal timeline view which gives a high level summary of the treatment. Finally some summary statistics, such as number of required days as an inpatient or total clinic visits, can be viewed.
The final area of functionality and access is the patient access. This will be done via the patient portal where a new link will be available on the portal to launch our application. Again user context and authorization will be passed to the application via tokens. Our patient personas are Jimmy, the cancer patient, and Susan, his mother. In this mode of access Jimmy and Susan can view the calendar or timeline views of any treatment calendars created for them by their providers. They can add individual events to the timeline, such as birthdays or vacations, to help them consider the impact of treatment on those important events. They can also comment on the treatment schedule which will send a notification to their provider; this could be a request to reschedule or a question on what will happen for a particular visit. Here Jimmy or Susan could also request to download .ical files for treatment appointments to integrate the events into their own calendar application. The purpose of the calendars currently created for patients at SCH is to help them with adherence. While we are working to help with decision support we do not want to negatively impact this important purpose of the calendar. Making the calendar digital and able to integrate with current calendar apps which Jimmy or Susan may currently use will hopefully make it even easier to follow the treatment plan. A key point to this feature area is that the application works well on mobile devices.

The next iteration of design and implementation will focus on to two more areas of functionality: patient’s (and their family’s) values and goals and genetic variant information. Users will be able to assign priorities to various values and goals along with open ended prompts and free entry which will allow capturing more in depth information about a patient’s wishes. This information which will be used to facilitate communication between all stakeholders and enable users to map values and goals to treatment plans. The mapping will then facilitate communication of how the treatment plan is personalized to the patient and help explain the recommended treatment plan. Reprioritizing, removing, or adding values and goals shall be reflected in the mapping immediately.

Currently, mapping genomic data to specific treatment options has limited use in pediatric oncology. This is a consequence of a limited number of targeted therapies currently available for use in the pediatric realm. While targeted therapies may be utilized in the treatment of adult oncology patients that have specific genetic variants in certain cancers, these drugs are often not yet approved for use in pediatric patients. However, if the patient has a relevant genetic variant which affects treatment decisions, information regarding targeted therapies specific for that individual patient’s genetic variant will be displayed. The type of variant and which drug or regimen it maps to will be shown and explained.

As mentioned we plan to integrate with the SCH EMR system which is Cerner. Some basic patient information will need to be retrieved from Cerner and if possible we would store the treatment calendar in Cerner so that a separate database which contains patient medical data would not be required. A key challenge for the project will be successful integration with Cerner. The planned approach is to use Fast Healthcare Interoperability Resources (FHIR), which is a medical data and web services standard, and the OAuth 2.0 (which conforms to the SMART on FHIR recommendations) authentication standard. Integration with Cerner and the authorization system also allows us to leverage the patient portal when authenticating patients to the application. While Cerner as a product does support FHIR16 more details on how much support SCH has implemented for FHIR and whether or not they allow data to be pushed back into Cerner still needs to be investigated. Building the application using the FHIR standard also increases the ability of the application to be deployed at another location.

Other key challenges in designing our tool include: 1) not interrupting workflow for key stakeholders (e.g. oncology providers, patients, caregivers); and 2) demonstrating sensitivity to each patient’s unique situation, preferences, and needs. Our tool will be made specifically to reduce the work required for oncology providers to prepare for treatment decision-making discussions. From a patient-centered perspective, our approach must remain sensitive to the concerns and environment of patients and their families. Due to privacy considerations, we have not been able to interview patients and caregivers. We have relied primarily on clinicians to serve as imperfect surrogates for the needs of their patients and caregivers.

It has been shown that determining the usability of applications implemented in a clinical setting is crucial for successful implementation11. Previous research has also shown that using multiple methods is more powerful than using a single method alone11. Thus we will use multiple methods for the evaluation of our tool; see Supplementary Material for a detailed description of our evaluation methods for each iteration of our design. We plan to utilize usability testing as we progress through each subsequent phase of development, consulting with providers,
caregivers, and patients as current functionalities are modified and new functionalities are added to ensure that our tool meets the needs of the end-users. In addition to usability studies, we will utilize online surveys to evaluate our proposed design. Additional details of the methods used for conducting surveys is provided in Supplementary Material. Surveys will be distributed at varying time intervals following implementation of our proposed design (e.g. 1 month, 3 months, 6 months, 1 year). We anticipate that data from these surveys will provide insight into how well our proposed design fits the needs of providers as well as elucidating areas for improvement.

We will use a number of different criteria in the evaluation of our proposed design. These criteria include: 1) ease of use; 2) amount of time required to create or customize treatment calendar; 3) missing function(s); 4) point(s) of confusion for user(s); 5) successful implementation. See Supplementary Material for a detailed description of how we will evaluate the effectiveness of our proposed design. Finally, successful implementation will be indicated by responses to the online surveys that users are still using our proposed design several months/years following initial implementation.

Alternative Solutions and/or Features Not Included
Originally we considered a solution focused on assisting providers with conveying genetic information to patients. We envisioned a tool to match patients and treatments based on genetics. However, in many pediatric oncology cases genetics is not the primary deciding factor when deciding on the best treatment for a particular patient. Rather, patient values play a huge role in the choice of treatment, especially in the case of relapse. After our interviews, we decided to go with a patient-facing application that would facilitate patient-provider communication helping patients and providers navigate this challenging time.

One theme commonly expressed by providers during our interview was that patients and their families would ask what treatment a physician would choose if they had to make the choice for their own child. This idea brought up the thought of a feature that showed statistics of prior patients, the choices they had made, and their outcomes. The feedback we received was that it may be coercive thereby negating the goal of developing personalized oncology.

Another area of functionality that would be helpful for providers and patients would be to help them discover other relevant treatments options, usually available clinical trials. The idea we proposed would be to automatically search ClinicalTrials.gov and display a list of trials for which the patient is eligible. Providers thought that it would be an excellent feature if it could be relied on to find an accurate list of trials for a particular patient. We decided not to pursue this functionality because of the technical challenge and the fact that it seemed a more provider centric feature rather than a patient focused feature.

Conclusion
Precision medicine is more than simply matching a patient’s medical data with a diagnosis and treatment. It encompasses that individual’s environment, family, lifestyle, goals, and values. While many healthcare providers are skilled at creating personalized treatment programs which incorporate all of these different elements, it can be very challenging and time-consuming to elicit the necessary information from a patient and then effectively communicate the treatment options and how they fit that individual. In interviewing pediatric oncology providers, we learned about the individualized treatment calendars that were being created and the value that patients and providers gained from them. A dynamic treatment calendar integrated with the electronic health record could be an effective communication tool, engaging patients more in the shared decision-making process. A tool which elicits a patient’s goals and priorities and matches them to treatment options can facilitate conversations with providers about what a patient wants. An electronic calendar makes it easier to share, update and integrate with a patient’s personal calendar for adherence once treatment begins. Our application integrates with the EHR and provides templates for calendar creation in order to fit into the provider’s workflow and reduce redundant work. In conclusion, CareMax enables precision medicine in pediatric oncology by personalizing cancer treatment recommendations toward the values, goals, and wishes of patients and their caregivers.
References


Abstract
Chronic diseases are the leading cause of disability and death in developed nations. Support for precision medicine in these complex disorders is needed. We propose a self-experimentation approach (SEA) as a mobile game to support precision medicine in chronic disease states. The SEA will be embedded within a learning health system to enable patient empowerment, healthcare provision, and medical knowledge generation.

Introduction
Specific Challenge: Chronic diseases are the leading cause of disability and death in the United States, where half of adults suffer from at least 1 chronic condition (1,2). These conditions are often complex with multiple factors (e.g., exposures, lifestyle, genetics) contributing to treatment response, morbidity risk, and disease prognosis. Given the complexities of chronic disease and resulting therapies, determining an effective, individualized treatment can take time to attain (if attainable). Herein lies an opportunity for precision medicine to advance chronic disease management and improve healthcare for millions of patients nationwide.

Current practice, for chronic disease management, is based on short patient-clinician interactions, where patient symptoms are discussed and clinician expertise is leveraged to direct a patient toward effective treatment and eventually, improved health. Clinicians are given the daunting task of aggregating and processing large amounts of information (e.g., blood work, diagnostic tests, prior conditions, family history, or knowledge about exposures) in a brief patient encounter to develop a treatment plan. This plan may include pharmacotherapy and lifestyle modifications where a patient is responsible to track his or her symptoms in response to treatment. At the next encounter, the treatment plan may be altered based on response to therapy. Personalization of treatment for chronic conditions is achieved slowly (often over months to years) and patients frequently fall between guidelines. Moreover, the time between patient-clinician interactions determines the success or failure of a prescribed therapy leaving much to be desired in current practice.

Several barriers limit precision medicine techniques from providing effective and efficient chronic disease management. First, many patients have difficulty tracking symptoms and communicating changes to clinicians. Some patients may mitigate symptoms while others may exaggerate the situation. Second, the current system depends on patient trust and clinician empathy to be conveyed during brief encounters. Third, treatment plans depend heavily on the clinician knowledge of the disorder and understanding of the patient context and preferences. Finally, accurate and through patient monitoring between encounters, which may be invaluable for treatment modifications, is lost or unreliable for decision-making in a brief patient-clinician encounter. In sum, improved support for personalizing medicine for complex and chronic disorders is needed.

We propose a self-experimentation approach (SEA) nested in a learning health system to support precision medicine for chronic disease states. This system takes input from the patient and clinician alongside other sources to fine-tune treatment through self-experimentation. The SEA will be incorporated within a learning health system to enable patient empowerment, healthcare provision, and medical knowledge generation.

Proposed Solution
Rational: SEA is based on the single-study design approach, where the study subject acts as both the intervention and control (3–5). This approach works well for personalized medicine as experimentation results are optimized to the individual rather than averaged across the population (4,6,7). Karkar et al developed a framework that applied self-experimentation for personalized healthcare with a mobile application for Irritable Bowel Syndrome (3). The results of this study demonstrate the feasibility and potential of this approach in applying personalized medicine to complex disorders. However, self-experimentation has not been fully leveraged to serve personalized medicine in either the scale of people or the number of medical problems (3).
The growing rate of mobile device ownership (68% of US adults) (8) and availability of medical apps for chronic diseases (9,10) make SEA feasible. Specifically, a gaming approach within the proposed framework may greatly support patients’ acceptance and adherence as gaming in health shows promising results (11,12). To ensure sustainability and dissemination of results, SEA should be nested with a learning health system, especially with recent advances in health information management and exchange (13). This integrated and gaming approach has more benefits than isolated apps as most only report measures rather than understand the patient’s unique characteristics and impact on her or his health in continues and robust scientific bases.

System Architecture: SEA has 2 main functions that are supported by multiple stakeholders:
1. Understanding patient characteristics to recommend optimal therapy and behavioral changes
2. Pooling and analyzing of aggregate data in order to update the current medical knowledge

Figure 1 depicts the proposed system with 3 integrated cycles: 1) patient empowerment, 2) healthcare provision, and 3) medical knowledge generation, described below.

![System Architecture of the SEA and learning health system.](image)

**Patient empowerment:** The center of the diagram has the patient and SEA as a game for a specific health condition. First the patient provides input about his or her condition and preferences and the SEA generates the first experiment. Additional input may include EHR, environmental data, wearable devices, or genetic data. Based on these input the SEA then guides the patient through experiment steps and data input as required. In cases where the patient cannot make regular data input, an additional person can provide input. This may be realized in a form of a game where the patient’s avatar has a hypothesis based-mission. As the avatar accomplishes the mission the patient receives feedback. Depending of the patients’ response to the experiment, the SEA provides other missions. On a
continuous basis the generated and collected data can be incorporated into the EHR. The patient empowerment cycle engages the patient by giving him or her responsibility and control over personalizing his or her own treatment.

**Healthcare provision:** The clinician may decide to make adjustments in the SEA to optimally fit the patient’s health condition and to avoid health risks. Data collected and generated by the SEA incorporated into the EHR may alert the clinician of certain risks (e.g., suicidal risk). The healthcare provision cycle ensures proper medical guidance and follow-up.

**Knowledge generation:** Results from many patients can be de-identified and pooled together to inform medical research in disease specific analysis pipelines. By leveraging predictive analytics, various hypotheses can be generated to refine the SEA. In addition, pooled results are available for the research community for further analysis. In both cases, the medical knowledge generated can be used to inform new or refine existing medical guidelines. These guidelines are then fed into the internal logic of the SEA along with EHR and clinical decision support modules. This process represents the knowledge generation cycle by using aggregate results to inform medical guidelines for a specific condition and patient demographic.

**Data processing and analyses**

**Individual level:** The single-subject study design is used to test the effect of the potential trigger and intervention in one experiment. Minimal of 3 data points during each of the intervention is necessary. Visual analysis, the most common analysis method for single-subject study (5) will be used to analyze the data within a phase. Changes in level, slope and trend will be calculated to describe the stability, variability, and directions of the changes between different phases. Additionally, two-standard deviation band method will be used to assess the variability within each phase. For time-series data, C statistic will be used to estimate the trends of the data.

**Population level:** The built-in data analysis model will be able to handle large scale aggregated data. Based on the population data, we will train models using machine learning and statistical methods to study interesting hypotheses. Phenotyping algorithms will be implemented to identify cohort with certain traits to facilitate clinical research study in the research community. In the meantime, the system provide data export function for users to export the data for their own data analysis.

All information in the SEA will be secured using the best available security protocols and data encryptions. HIPAA and GINA regulations will be considered for patient privacy and de-identification when aggregate-level data are retrieved. As the SEA can be used by subjects located in different states or countries, security regulations will need to be customized. This customization will be guided by a panel of ethical experts and stakeholders in the territories where the SEA is used to ensure ethical and legal adherence.

**Use Case – Major Depressive Disorder (MDD):** A promising application for the SEA is MDD, a complex phenotype affecting 6.7% of adults in the United States (14). MDD is characterized by loss of interest in usual activities or depressed mood accompanied by fatigue, difficulty concentrating, changes in appetite or sleep, feelings of worthlessness, or suicidal thoughts lasting 2 or more weeks (15). Often a chronic illness, MDD can lead to negative outcomes including decreased quality of life, lack of ability to function, negative impact on relationships and work capacity, or even suicide (16).

Finding the right treatment plan for an individual patient with MDD takes fine-tuning and often trial and error. Although over 100 MDD apps exist, providing mood and activity tracking, these apps stand-alone and do not provide personalized interventions based on the patient’s unique condition. MDD treatment generally consists of pharmacotherapy, psychotherapy, and lifestyle modifications (15). The first two categories of treatment (pharmacotherapy and psychotherapy) require psychiatrist or psychologist oversight. Here the SEA can identify optimal treatment based on patient response. In addition, the SEA can provide plans for lifestyle modifications. These functionalities are described below.

**MDD self-experimentation:** Multiple lifestyle modifications are recommended to MDD patients including sleep hygiene, physical activity, diet, mindfulness, or light therapy (15). The most effective lifestyle modifications are the ones that a patient will consistently follow (15). The SEA can effectively identify the optimal lifestyle modification. Through the SEA, a clinician can enter recommendations for changes in lifestyle such as amount of exercise each day, increasing mindfulness, changes to diet (e.g., increasing omega-3 fatty acids or folate), or light therapy. The patient could then input personal preferences about type of activities and duration. The SEA would randomly assign different lifestyle modifications to the patient. A mood and functioning assessment would then be required at the start, halfway through, and at the end of a lifestyle modification plan. Each day the SEA would provide a lifestyle task that falls
within the suggested routine and ask for the patient indicate completion of the task. Taking into account adherence, mood, and functioning results, the app would either suggest a different routine, or provide a similar routine. Once a routine that works well is found, the app continues to provide these suggestions until a decline is noticed. The SEA can also provide clear monitoring of how often a patient is doing with lifestyle modifications, and the gaming aspect can help provide motivation to adhere to the program.

**MDD monitoring:** Currently, mood and how well a patient is functioning is mainly assessed at clinical visits. The SEA can provide a more in-depth picture of a patient’s response to treatment by assessing mood and functionality more regularly. Mood can be assessed through a series of questions within the app. Functionality can be assessed by randomly asking the patient what he or she is doing throughout the day and connecting with wearable technology (e.g. Fitbit).

Results of the self-experimentation alongside mood and functionality assessment will be summarized and reported to the patient and clinician. These data can be used to further inform treatment recommendations in order to find a plan that works for the patient. In addition, results from the experimentation can be fed back into the research community and inform future research by providing details about the patient’s demographics (e.g., body mass index, prior lifestyle, current medications). In this way the SEA aids personalization of treatment for an individual patient and informs treatment of MDD for similar patients.

**Alternative Solutions:** Several additional solutions were considered for SEA for MDD. First, another option could be support for clinician guided experimentation with pharmacotherapy. If first line therapies are not effective it can be difficult to determine which additional option is best. Further research is needed to develop guidelines for second-line pharmacotherapy for MDD, and the SEA could provide this experimentation. Consultation with practicing psychiatrists, however, led us to lean away from this level of experimentation as clinicians may be less likely to use this functionality. Second, we considered supporting psychotherapy experimentation. This layer of treatment often involves developing very specific care plans. Here SEA could provide evidence for what psychotherapy care plans are effective for specific patients. Therapy plans, however, are often very specific to the patient’s symptoms, which may make it difficult to develop general MDD algorithms. Finally, a third alternative solution is targeting the SEA to sub-clinical, or pre-diagnosis patients. Here the app would focus entirely on lifestyle modifications that may help the patient improve symptoms enough to not require medical care. If user symptoms decline, the SEA could recommend the patient seek medical care. This final solution is feasible and could be implemented by making the app available to the general population.

**Implementation and Dissemination**

**Gaming and alternative approaches:** According to the Entertainment Software Association, 63% of households in the U.S. include at least one regular game player, 47% of the players are between 18 and 49 years old. Gamer gender distribution shows that 41% of the players are female (average 44 years old) and 59% are males (average 35 years old) games (17,18). Clearly, games are widely used and accessible across demographics. The SEA leveraged as a game can potentially increase acceptance and usability of the app.

Although games are appealing to many, some individuals may not prefer games, as a result, an alternative representation will be provided that instead uses forms and reports. Following principles of effective system design these reports will incorporate familiar symbols and metaphors (e.g., to-do lists, calendars and reminders) (19).

**User interface design process:** Interviews and focus groups will be held with clinicians and patients to determine what functionalities are desired for a specific use case. Card sorting activities and rapid prototyping with end-users will be employed to develop the patient, clinician, and researcher facing user-interface for the SEA.

**Information exchange:** As the application is required to collect and provide data from/to many separate entities, Health Information Exchange (HIE) will be intensive. To ensure proper syntactic and semantic interoperability the application data representation should follow the commonly used HIE standards such as SNOMED-CT and LOINC. Communication between the EHR and SEA app will use available messaging standards such as HL7-FHIR. In addition, SMART on FHIR create an integral platform that allows data communication among portable healthcare application and the underlying EHRs (20,21).

**Cloud-based data architecture:** The SEA system will use the cloud-based data architecture to store, process, and transmit data. HIPAA protected private (e.g., Center for High Performance Computing at the University of Utah) and public (e.g., Amazon Web Service) clouds will be utilized as SEA service providers.

**Evaluation Plan**
Fully developed versions of the SEA will be evaluated for usability and efficacy. Usability would be assessed based on survey evaluation by end users including clinicians, patients, and researchers working with the phenotype of interest. Efficacy of the SEA for a phenotype of interest would be evaluated by a randomized controlled trial. In this evaluation, patients with the phenotype of interest would be randomly assigned to a treatment plan that uses the SEA or standard clinical care. Comparison of symptoms (patient mood and functionality) would be assessed at initial diagnosis and every month until treatment is no longer needed or for one year. Quality of life (based on mood and functionality assessment) and time to improved symptoms would be compared between the two groups to evaluate the efficacy of the SEA.

**Strengths and Weaknesses**

Our proposed solution has the strength of integrating mobile application with the larger cycle of learning health system. This ensures the benefit for both individual and population level in addition to updating the medical knowledge through providing valuable and structure pool of data. Also, it makes use of the currently available HIE standards to ensure semantic interoperability and avoid redundant or manual data entry as much as possible. The use of the gaming approach supports system acceptance and usability. In addition, this approach fosters patient empowerment and engagement in healthcare process which is considered as the “Blockbuster Drug of the Century” with the aim to improve quality, decrease costs and mitigate disparities (22,23). Here the patient is responsible to understand his or her health condition and participate in finding the optimal treatments or recommended behavioral changes. In conclusion, both the use of games and the integration within a larger learning health system may be considered a valuable addition over the originally developed self-experimentation framework by Karkar et al (3).

The proposed framework still faces some challenges, however. First, the ethical, legal and regulatory framework need more collaborations and discussions among the involved stakeholders. Second, finding HIE standards that capture all concepts may be a challenge and some additional new concepts may need to be proposed to the corresponding standard development organization. Finally, structuring an idea for a game that both fits the SEA and is interesting for the intended population is not a straight-forward task. A thorough review of the currently available games and engagement of intended users and game experts are necessary to reach a useful and usable self-experimentation game.

**Conclusion**

The SEA is an innovative approach to help patients with chronic conditions find therapy that improves their unique conditions. The patient empowerment cycle enables users to take responsibility and control over their conditions. The physician provision cycle improves communication between patient and clinician to ensure shared decision making. The knowledge generation cycle informs future treatment of chronic conditions. This unique integration of SEA with a learning health system can support precision medicine for complex, chronic conditions—potentially improving quality of life for millions of Americans.

**References**

2. Fifty Years of Progress in Chronic Disease Epidemiology and Control [Internet]. [cited 2016 Jun 30]. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/su6004a12.htm


Breast Cancer Personal Trajectory Tool
Rebecca A. Marmor, M.D., Elizabeth Epstein, B.A., Meera Reghunathan, B.S., Mitchell Boldin, M.S.
University of California San Diego, San Diego, CA

Abstract

Patients diagnosed with breast cancer face a myriad of treatment decisions. Although there are a variety of resources to help patients make decisions, each is limited in its scope and ability to offer personalized information to users. We propose a personalized decision-making tool that synthesizes data from large trials, granular data from the UCSD Clinical Data Warehouse and qualitative data obtained from online health communities. Information about treatment options and most likely outcomes, will be presented to patients through the lens of their individualized risk factors and prior treatment decisions. We outline our iterative three-phase design process which concludes with a feasibility study.

Introduction

Upon diagnosis of breast cancer, many patients face a variety of treatment decisions, some of which may have significant physical and psychosocial implications.\textsuperscript{1,2} Decision-making in breast cancer is commonly patient-driven. Patients are often required to select from a menu of surgical procedures which include: lumpectomy, mastectomy, contralateral prophylactic mastectomy and/or reconstruction. They might be offered radiation and/or chemotherapy, and asked to take estrogen modulating medications.\textsuperscript{3,4} Patients seek out second opinions and can learn that providers have different recommendations.\textsuperscript{5} They also seek information from peers and online health communities (OHCs).\textsuperscript{6–8} The volume of information can be overwhelming, confusing and sometimes even conflicting. It can be very difficult for patients to predict what their own most likely outcomes are, given their individual risk profile comprised of their current health status, genomic markers and family history. Additionally, past treatment decisions may influence future risk of disease recurrence and metastasis; this knowledge may inform patients’ current decisions.

Figure 1. Sources of Information for Breast Cancer Patients

Information for Breast Cancer Patients

There are three types of information that breast cancer patients commonly use to make treatment decisions: clinical data based on large research trials, anecdotal data from friends and family and data obtained from consultation with health providers (Figure 1). Data from large research trials may provide patients with estimates of risk of local cancer recurrence and overall survival depending on the treatment options they choose. This data is helpful because it can provide long-term follow-up of patients, and is based on large numbers of patients. However, it is limited by its lack of granularity, and may not address more specific questions patients have regarding treatment options.\textsuperscript{9} Also, it may not account for an individual’s specific risk profile. Secondly, patients may consult peers regarding various treatment options. Patients may also seek out similar advice on OHCs.\textsuperscript{10,11} Narratives shared by other patients are often most influential in informing patients’ understanding of how treatments might promote or contradict their idea of health.\textsuperscript{8,12} Numerous studies have explored the benefits of engaging with fellow patients via OHCs about their personal experiences when faced with the same disease.\textsuperscript{13–16} Benefits such as hearing how someone else has coped with the same illness and managed treatment day-to-day provide context and help patients understand their disease beyond medical explanations.\textsuperscript{11} However, patients can have a poor understanding of how their own disease process...
differs from their peers’. As our understanding of cancer is rapidly evolving, with increased emphasis placed on cancer genomics, patients might not be aware of the subtleties of their own disease that affect their prognosis. Similarly, a verbose member of an online health community may complain about her cosmetic outcome following surgery, and create a large volume of posts expressing these views, even though vast majority of community members are satisfied with their outcomes, thus skewing other users’ perceptions of what the most likely outcome is. Finally, information and counseling obtained from health providers may be tailored for an individual’s risk profile and include detailed information about the patient’s current health status, genomic markers and family history. However, information from clinicians might be seen by some patients as untrustworthy, with some patients ultimately preferring to seek out their own information. Information obtained from clinicians also lacks the subjective experience provided by peers and thus might be perceived as less reliable.

Decision Aids
Prior research has been conducted to develop decision aids (DAs) for breast cancer patients. There are several studies confirming the utility of such DAs in decreasing decisional conflict and increasing knowledge and satisfaction. Although providing helpful preliminary data, these decision aids are limited in scope- most address only one treatment decision (e.g. mastectomy v. lumpectomy). None currently support the patient throughout the entire disease course. Additionally, many of these prior DAs have not been updated to reflect recent changes in practice, such as new chemotherapy regimens which may affect outcomes. Perhaps most importantly, no DA currently incorporates information about patient’s subjective experiences in a systematic way, despite the emphasis patients may place on this information when making decisions.

Statement of Problem
Breast cancer patients currently face a wide range of treatment options for their disease. Although there is a plethora of resources available to help them choose among options, these resources are currently limited in their ability to offer personalized information about recurrence risk and survival benefit, based on the patient’s own demographics, health history and tumor characteristics. Additionally, these resources are currently disjointed, addressing a variety of separate concerns and may even sometimes be conflicting. We propose a DA which would incorporate several different types of information, filtered through the lens of a patients’ own unique risk factors, to help patients make treatment decisions.

Alternative Solutions and Designs
Since there are a plethora of resources available for breast cancer patients already, we have carefully considered alternative solutions to this problem. First, we considered merging and updating the content of existing decision-making aids, into a unified aid to support patients throughout their disease course. This solution offered the benefit that many of the pre-existing aids have already undergone validation testing, and have demonstrated efficacy. However, we would still need to repeat validation of this new aid. This solution would also not allow us to incorporate additional data sources, which we believe are vital to decision-making. We also considered working with existing online health communities to refine search capabilities, so that patients could seek out specific treatments and concerns regarding these treatments. This approach offered the advantage of pre-existing large, active and engaged memberships. While this approach may have allowed us to show the frequencies of certain outcomes among community members, it would have not allowed us to superimpose clinical data. This approach is also limited in its ability to provide users with personalized information, tailored to their own risk profile. Even if the search function was refined to such a degree that patients could perform sophisticated searches to answer specific questions, it would be limited by patients’ abilities to recognize users’ disease as similar to their own.

Solution Design and Development

Design Process
We plan to use an iterative, patient-centered design process. The process consists of three stages: 1) Design Requirement Extraction 2) Iterative design and prototype development 3) Evaluation of Trajectory Explorer. We will devise an initial design based on the information obtained in stage one, which will be updated based on feedback obtained throughout the stages from users (Figure 2).
Stage 1: Design Requirement Extraction. We plan to begin our design process by performing in-depth qualitative interviews of breast cancer patients at all stages of treatment. The purpose of these initial interviews is three-fold:  
1) To gain insight into the decision-making process of patients not available in the literature. Specifically, we are interested in what types of decisions were especially difficult for patients to make, what resources they consulted with in order to make those decisions and what they liked/disliked about those resources. For patients who have already made decisions, we will ask them “what they wish they had known,” and for those who are presently in the decision-making process, we will ask them what options they are considering and what information they would like to have available.  
2) To learn about the preferred methods of patients for receiving information. We will ask patients about their comfort level using mobile phones, apps and computers. We will ask about their access to the internet. Finally, we will ask about their comfort interpreting basic statistics including probabilities.  
3) To gain feedback about mockups of the proposed solution, so that we can make changes to improve our design.

We also plan to consult with breast cancer clinicians, to gain insight into the perceptions of providers about the most difficult treatment decisions for patients, and what information they wish they were able to provide patients with that they cannot currently. This part of the process will also include an in-depth literature review of decision-making among breast cancer patients, and prior DAs. We will also conduct brainstorming sessions. At the conclusion of this phase, we will have formulated our design requirements.

Stage 2: Iterative Design and Prototype Development. Our proposed solution is a mobile application (app) that assists in patient decision-making. The app incorporates three distinct data sources for patients, and is readily used by patients during consultations to keep track of information relayed by clinicians. This solution will provide personalized information to users, based on information entered on a personal profile. The app will help users learn about different treatment options for breast cancer, expected benefits of the treatments, given their personal risk factors, and the experiences of other similar patients who have selected the treatments (Appendix A). This information will help patients weigh competing treatment options, and eventually choose one. Finally, it will serve as a repository for users to store additional information they have learned from other sources, including clinicians. We plan to develop a web application with mobile users in mind. A web-based app will have the benefit of being accessible to both iOS and Android users. Moreover, although we believe that users will primarily access the app via smartphone, a web-based app will allow users to access the application via their computer if they do not have a smartphone or if they prefer not to use a smartphone.

Data Components

The three layers of data (Appendix A) we plan to include in our model are:  
1) Clinical Data from large research studies: This will be obtained through literature review, and based on current NCCN guidelines.23 This data will primarily describe risk of local recurrence, distant metastasis and death. When available, we will include individualized risk factors, including BRCA status, age, race and tumor characteristics, to personalize the
information patients receive. This information is limited in granularity, but offers the benefits of large numbers of patients and long follow-up. 2) UCSD Clinical Warehouse Data. Using data from the 8,000 breast cancer patients seen at our local center, we will perform predictive modeling to identify the most likely outcomes for breast cancer patients based on individualized treatment decisions. This data is more granular than that provided by large studies, and may provide information of interest to patients (e.g. what are the chances that I will not be able to complete my course of Tamoxifen? What are the chances I will need an additional surgery to get complete clearance of the tumor following lumpectomy?) The questions we will answer with this data will be largely informed by our patient interviews, and will seek to answer those questions of greatest interest to patients. This data set is smaller than national data, but will allow us to help patients answer questions that may be most important to them. 3) Qualitative Data obtained from online health communities. We have scraped nearly 320,000 posts from two major OHCs (WebMD cancer community, American Cancer Society forum) to assess patients’ questions and needs concerning choosing breast cancer treatments online. We will use these posts to extract patient experiences in quotes and sentiments around treatment options. We will extract breast cancer related terms, the sentiment around the term, and quotes using natural language processing techniques and tools such as Unified Medical Language System24 and Linguistic Inquiry Word Count.25 Additionally, we will perform a keyword search regarding specific patient concerns about various treatments (e.g. burns and radiation, chemotherapy and fertility). We will extract this data, along with demographic information of posters, to provide patients with subjective information from similar patients (e.g. by age, cancer stage, BRCA status, etc.) about cancer treatments. Moreover, users of the app will be able to post their own experiences for other users to view, which will provide additional information for patients. Users will be able to rate posts as helpful so that the most valuable and informative posts are shown preferentially to end users. We will verify the validity and reliability of all data with experienced breast cancer clinicians at the UCSD Comprehensive Breast Cancer Center.

Key Features

In order for the app to educate patients, users will need to be able to engage with it in meaningful ways. The following features will support this engagement. 1) User Profile (Appendix B). Creation of a profile is very important, to ensure customization of information delivered to app users. Users will be invited to share as much information as possible about their health and disease status. However, because additional information about a patient’s stage, tumor characteristics and genetic markers may not be available initially, and the patient’s risk of future cancer-related events will change depending on the past treatment options selected, the profile will serve as an easily accessible centerpiece of the app. Depending on their location on the treatment trajectory, users will be provided with appropriately timed prompts to update their profiles (e.g. a patient who has undergone surgery will be prompted to update her stage and pathology results two weeks after her surgery date). This profile will be the basis of our collaborative filtering system, to ensure that the user is provided with information about similar patients.26 2) Customized Information Displays (Appendix C). Users will be able to customize their displays to see information from any one, or all data sets (e.g. large trials, data warehouse and/or qualitative). This feature will allow them to receive the information they are most interested in. It will also educate patients about the benefits and limitations of each data set, with suggestions about how the given data set may be interpreted. 3) Annotation Mode. To prepare for upcoming visits with providers, patients will be able to compile lists of questions they hope to have answered. Depending on the specialty of the provider, the position of the patient on the treatment trajectory and the patient’s individual health characteristics, the app will suggest appropriate questions for the visit. The patient will be able to add additional questions to the list. At the time of the visit, the patient will be able to take notes with the answers. Additionally, the patient may be able to record the clinician’s answers to the questions and store this in the app.

We will low fidelity mockups of the app with a tool called NinjaMock, which allows for the rapid production of functional and adaptable mockups.27 These affordances are especially important in our iterative design process, because certain characteristics of our solution are subject to change. We will seek out user feedback after each iteration, conducting approximately three interviews and then improving on our design. We will plan to conduct 20-30 interviews total. When we are not receiving any further significant suggestions from users, we will develop a high fidelity, interactive prototype in order to get more detailed user feedback. This design stage will end with usability testing for the final prototype.

Stage 3: Evaluation of Trajectory Explorer. As we will be using an iterative design process, we will be seeking patient and clinician feedback throughout development. Evaluation of our tool will involve a three-phase feasibility study. Formal evaluation of the aid will commence once we have a final working prototype. We will use O’Connor’s Decisional Conflict Scale (DCS), which has been used in multiple previous breast oncologic studies to gauge decisional conflict among patients28. The DCS is composed of 16 items, which assess 5 different subscales:
(1) “Informed”, which evaluates awareness of options, benefits, and risks, (2) “Values Clarity”, which evaluates feelings about benefits and risks, (3) “Support”, which evaluates feelings of support and pressure from others, (4) “Uncertainty”, which evaluates the participant’s confidence, and (5) “Effective Decision”, which measures the likelihood of adhering to the decision.

We propose to evaluate our decision aid using a three-phase approach: Phase I: We will conduct a small pilot study of 30 newly diagnosed patients at our institution, administering our decision aid to all of them one week prior to their first clinic visit. We will assess their response with the DCS and qualitative interviews to ascertain their reactions to the aid, as well as assess how it may be improved. We will incorporate these recommendations into the next version of the tool. Phase II: This phase will incorporate revisions to the aid made based on the results of our phase I study, and include 50-75 newly diagnosed patients at our institution, all of whom will receive the decision aid one week prior to their first clinic visit. We will again administer the DCS, at various points (immediately following their first clinic visit, just prior to surgery, at their first postop clinic visit) along with the FACT-B and EORTC QLQ-30 two validated instruments used assess quality of life among cancer patients. Finally, we will include a survey for patients, to obtain recommendations about how to improve the decision aid. Phase III: This will be a large, randomized, single-blind multi-institutional study to evaluate the efficacy of our DA. Newly diagnosed patients will be randomized to receive either our DA or a traditional DA one week prior to their first clinic visit. Providers will not know which arm of the study patients are in. Again, patients will be assessed with the DCS, FACT-B, EORTC QLQ-30 and a survey to assess their satisfaction with the decision aid. Patients will have access to the tool for one year, and will be assessed with the aforementioned instruments at routine intervals.

Conclusion

Breast cancer is a disease uniquely positioned to benefit from the advances of personalized medicine. Patients, who are often asked to choose from a variety of treatment options, may consult a variety of conflicting and disjointed resources in order to make informed decisions throughout the disease course. We propose the development of a decision aid application, utilizing data from large trials, a single institutional data warehouse and subjective data from OHCs to help patients make informed decisions. This information will be delivered through the lens of the individual patient’s personal risk factors and prior treatments.

References


Improvement of Patient Reported Depression Outcomes using EMR Integrated Genomics and Mobile Assessment

Anjana Ramnath, MS¹, Timothy A. Green¹³, Diana E. Rickard, MHA, MSHI¹², Lincoln R. Sheets, MD

¹MU Informatics Institute, University of Missouri, Columbia, Missouri, USA; ²Department of Health Management and Informatics, University of Missouri, Columbia, Missouri, USA; ³The Tiger Institute for Health Innovation, Columbia, Missouri, USA.

Abstract

Mobile assessment of patient reported data related to depression, in combination with genomic SNP panels, can provide additional sources of data to help physicians make better treatment plans. In order to be effective as an intervention, these additional data must be integrated into the physician workflow. This study will leverage an existing ambulatory mood assessment app, MoodTrek™, and available knowledge related to single nucleotide polymorphisms that impact metabolism of common anti-depressants. Patient reported data will be combined with genomic data in an EMR integrated dashboard that physicians will use during normal treatment planning. A three stage study design will assess the acceptability of the patient facing mobile intervention, workflow impact to physician users of the EMR dashboard, and clinically relevant patient outcomes.

Introduction & Background

Depression is a chronic mental illness that affect millions of patients around the world. Prevalence in U.S. has been measured at 4.1% for major depression and 9.1% for any depression [14]. Major depression accounts for 3.7% percent of the Disability-Adjusted Life Years (DALYs) in the U.S [15]. Current recommendations for treatment of depression is a combination of psychotherapy and antidepressants. Primary care physicians are on the front lines for treatment of depression, treating approximately one third to one half of adults and almost two thirds of older adults with major depression [16]. Patients are commonly assessed in regards to the symptoms of depression such as quality of sleep and sleep patterns, and other subjective measures including appetite, concentration, energy levels, and suicidal ideations. Patients self report how they have been doing in these areas since the last clinical encounter. This information is then used to modify treatment, such as changes in prescriptions and intensity of care.

Nearly three-fourths of adolescents (age 13-17) in the United States (US) have access to a smartphone and 92% report going online daily or more frequently [7]. The rise of smartphone health applications are aimed to improve the retrieval of information in clinical practice, two-way communication between patient and medical provider, provide patient resources and education, and connect to external devices (e.g. sensors, monitors, DME) to seamlessly feed and push information [6]. By enhancing the data capture outside of discrete times (e.g. office visits) we strive to decrease the prominent problem of recall bias in the accuracy or completeness of the recollections over time with adolescent patients with depression. This proposal aims to develop, enhance, and implement an ecosystem of tools to improve adolescent depression functioning through the capture of personalized, patient reported data to support patient-centered decision support.

Mental disorders are complex and effectiveness of drugs usually used to treat common mental illnesses such as depression may be linked to genotypes of the particular patient [11]. To better understand the effects that different variants and mutations have on mental illnesses and the way patients metabolize drugs for the same, various SNPs associated with Cytochrome P 450 (CYP450) metabolism could be tested. CYP450 genes affect how the body metabolizes selective serotonin reuptake inhibitors (SSRI) drugs as well as many other drugs. CYP450 genotyping (genetic testing) is being proposed as a means of guiding the use of SSRI drugs to improve the effectiveness of treatment for depression.
In addition to the traditionally viewed precision medicine technique of genomics, patient reported outcomes and patient generated data are important classes of data to be considered for precision medicine. Collection of patient reported data via mHealth tools in support of depression management have been well studied. In particular, recent research into the use of mHealth tools for ambulatory assessment in the mental health space have shown strong support for physical activity monitoring and single item mood score collection. A UK based study published in 2015 [1] found that a single question ambulatory assessment of depression using text messaging was found to be a valid measure of latent affective depression as validated against the PHQ-9 [10]. A similar study from 2015 [2] conducted in the US also showed a correlation between daily and one week average mood scores, and the standard PHQ-9 assessment scores. Another 2015 study [3] found that use of mobile phone sensor data could be used to classify patients with (PHQ-9 score > 5) and without (PHQ-9 score <= 5) depressive symptoms, with an accuracy of 86.5%.

In addition to strong support for correlating simple single score mood score to validated measures such as PHQ-9, other studies in the mental health space have found positive results with regards to usability and adherence. FOCUS [4], a smartphone based system, provided pre-scheduled and on-demand real-time/real-place illness management in schizophrenia. 33 patients with schizophrenia used the tool for one month. 90% of participants found the tool as highly acceptable and usable. Despite the short duration of the study, reduction in psychotic and depressive symptoms was noted in the study cohort. Another mobile intervention used in a 2014 study [5] was personal digital assistant to prompt engagement in self-management. Preliminary results showed adherence similar to that reported in non-psychiatric populations, with high participant satisfaction. Based on these studies, mHealth apps and devices seem feasible and acceptable in augmenting psychosocial interventions for severe mental illness.

When combined with genomic data of the patient such as presence of pathogenic variants or mutations, these applications can provide a valuable source of patient reported data that can be used to tailor treatment for individual patients.

**System Design**

**MoodTrek™**

MoodTrek™ ([http://www.mood-trek.com/](http://www.mood-trek.com/)) is designed to minimize the impact of the recency effect on the diagnosis and treatment of depression. Composed of two parts, MoodTrek™ is a patient-facing app and a provider-facing view within the Electronic Medical Record (EMR). The app allows patients to track their moods, activity, and sleep, as well as journal about their experience. This information can be automatically shared with a physician within the EMR. By integrating this patient generated data into the EMR, MoodTrek™ helps streamline the physician workflow by providing the care team with all of the benefits of a traditional mood chart plus a richer data set powered by the patient’s actual activity in real time. These integrated tools allow caregivers to accurately understand their patients experience with depression. Rather than relying on anecdotal conversations related to how an individual “feels”, they are able to see the flow of mood over time compared with activity and sleep levels.

MoodTrek™ is available on both iOS and Android platforms, and can connect with and integrate data from the Fitbit activity tracker. When using a Fitbit with the app the patient’s activity and sleep information can be captured accurately and completely with next to no action on their part. Through an invitation process initiated by the physician, patients can control which data are shared with the physician. The patient has full control over the type of data that can be shared, including sleep data, steps, mood, and journal entries. By using a SMART app interface[9], the MoodTrek™ data can appear alongside other clinical data within any physician’s EMR that supports the interface, thus providing a seamless workflow experience in using the data as part of normal clinical workflow.
Current sequencing methods range from running a genomic panel, which is focused on a specific set of genes, to whole-genome sequencing. Establishing the analytical validity of a sequencing diagnostic, however, is considerably more difficult for whole-exome or whole-genome approaches, in which the average depth of coverage is much lower than for targeted gene sequencing of a defined panel of genes. The establishment and maintenance of the analytical validity of a diagnostic test requires substantial effort. In a clinical laboratory, this entails initial and periodic testing on reference standards to demonstrate the reproducibility of the assay [12]. In contrast, panels targeting specific single nucleotide polymorphisms (SNPs) are in use and have a clinical track record. The cost of clinical-grade whole exome sequencing (WES) is high, whereas the cost of running a sample on one of such panels is relatively less expensive. As such, for treatment of mental illness, we will focus on usage of genomic information in SNP panels in order to better understanding how patients metabolize various drugs, including the most commonly prescribed ones. Panels will be run to test patients on variants involved in CYP metabolism. Designing panels to test for specific CYP gene variants that are involved in processing the drugs used in depression is also cost effective. The panels ideally would cover the spectrum of pediatric and adult clinical genetics in mental illnesses.

EMR Dashboard

By combining the MoodTrek™ data with genomic data from SNP panels in a single dashboard, physicians will have ready access to all necessary data to make a personalized treatment plan or adjust existing therapy. In a single view, physicians will be able to see mood scores over time, step counts, sleep amount and quality, journal entries and a journal word cloud, and SNP panel data (Supplement). The intention of the dashboard is to provide all relevant information necessary to make a clinical decision on the efficacy of pharmacologic treatment and other interventions. Because the dashboard is integrated into the EMR, other, more classical forms of clinical data are available through the normal means of chart review.

Study Design

A three stage study design is planned to optimally evaluate the 1) usability of the patient-facing MoodTrek™ application, 2) usability of the EMR provider dashboard, and 3) a clinical trial to measure outcomes of use of both the MoodTrek™ application and the EMR provider dashboard.

In the first stage of this proposed study design, we will evaluate the acceptance of a new patient-facing application, MoodTrek™, designed to minimize the recency effect of the diagnosis and treatment of patients with depression. To identify potential barriers, this stage aims to evaluate the usability, specifically perceived usefulness and perceived ease-of-use, of the patient-facing MoodTrek™ application. By adopting the empirically validated Technology Acceptance Model (TAM), we will understand the factors influencing how and when users will access the new application (Davis, 1989). This study will recruit ten potential users who are within the target demographic age (13-17) and who self-report current or past history of depression. Enrolled participants will be asked to complete a 33 item online questionnaire, modified from Gagnon (2012), following a one hour block of uninterrupted and self-promoted interaction with the MoodTrek™ application. The online questionnaire measures 8 dimensions including: perceived usefulness, perceived ease-of-use, attitude, compatibility, subjective norm, facilitators, habit, and intention [8]. Results from the questionnaire will be analyzed using descriptive statistics to determine the mean, standard deviation, and correlations between the 33 variables across the 8 dimensions (Suppl 1).

The second stage of the proposed study design surrounds usability testing of the EMR provider dashboard. The System Usability Scale (SUS) [13] scale will be used to test ease of use for clinicians when navigating the EMR-Genomic dashboard. The trial period will be 24 weeks, entailing dashboard use throughout by the clinician to treat and adjust treatment as necessary.

The third stage of study design consists of a randomized clinical trial. In order to assess the clinical efficacy of the MoodTrek™ app on outcomes that are important to patients and clinicians, a randomized trial will be conducted using a matched intervention group of patients who will use the MoodTrek™ app, and a control group who will not use the MoodTrek™ app. The primary clinical outcome measure will be the PHQ-9 [10]. In addition, patient reported outcome measurements will be collected using the PROMIS Emotional Distress-Depression – Short Form 4a measure (suppl 2). Secondary outcomes to be investigated will include the frequency of patient encounters and medication prescription stability.

The trial period will be 24 weeks, with the initial consenting clinic visit serving as the kickoff visit, and a scheduled visit at the end of the 24 week period. During this time, patients will be followed as normal, including
normal follow up clinic visits as necessary. Both study arms will be asked to submit the PROMIS Emotional Distress-Depression – Short Form 4a measure monthly through an emailed survey. The PHQ-9 will be collected at the beginning and end of the study period for each arm by the clinician. Retrospective data pulls will be used to determine frequency of clinic visits during the study period, and EMR sourced data on prescriptions will be used to measure the stability of pharmacologic interventions over time.

Calculating for 80% statistical power, and a 10% attrition rate, 682 patients will be enrolled in the study. Target patients will include 13-17 year old adolescents with a diagnosis of major depressive disorder. Patients will be recruited in clinic, where supplementary materials can be shared and consenting can be performed with patients and patients’ guardians.

Implementation

The MoodTrek™ app and EMR integration for the clinical trial will be implemented at University of Missouri Health Care (MUHC), in the primary care setting. MUHC uses the Cerner EMR system. The clinical integration of Mood Trek with the Cerner platform has already been developed and is available to install in the MUHC system. The MoodTrek™ app is available on the Google Play store and the Apple Store. Training on the use of the MoodTrek™ integrated EMR component will be provided to all primary care physicians through training outreach performed by members of the research staff. Training for users of the MoodTrek™ app will be minimal and will consist of flyers handed to the study subjects during the consenting and enrollment process.

Conclusion

Mobile interventions such as MoodTrek™ promise to bring personalized data into the hands of the clinician. Combined with genomnic sources of precision medicine, such as metabolic panels, the described system is expected to deliver better care. We expect to see this reflected in improvements in overall PHQ-9 scores and improved responses to the PROMIS Emotional Distress-Depression – Short Form 4a measure. Additionally, we expect that when clinicians leverage these sources of precision data, the stability of pharmacologic treatments will improve, with fewer changes per patient in dosing or drug. Also of importance, given the shortage of adolescent psychiatry services, is the expectation that patients will use fewer services when patients and physicians are connected to more precise and objective forms of data outside of the clinic.

References


PrecMed: A collaborative environment for precision medicine research and practice

Jianlin Shi MS, MD, Lance Pflieger
Biomedical Informatics Department, University of Utah, Salt Lake City, UT

Abstract

The goal of precision medicine is to transform healthcare by providing patient-specific care through the application of evidence-based practices gained from large and often heterogeneous health-related data. However, current evidence and practices can be statistically insufficient to infer proper precision healthcare guidelines, especially in respect to common genetic variants. Hence, it is important to incorporate and utilize data from an array of disciplines and institutions. PrecMed is a platform designed to foster collaborations, not only for data use and generation, but also for crowd-sourcing of ideas and application development with an emphasis on peer evaluation. PrecMed also supports distributed data analyses to aid with large cohort studies and provides real-time results aggregation, facilitating iterative application development. Furthermore, it provides an easily accessible place for disseminating precision medicine research information, results and applications.

Background

Personalized medicine aims to break from traditional population based medicine by providing individualized treatment tailored specifically to a single patient. It is defined commonly as “The right treatment, for the right person, at the right time” and in many circumstances should include the right cost. While personalized medicine is associated frequently with advances in genomics, personalized medicine also incorporates clinical, biological, environmental, social, economic and life style factors. Ideally, the collection and analyses of these personalized factors will lead to predictive analytics to identify a person’s risk for developing disease, guidelines for prevention and the precise course of treatment. Realistically, multiple barriers need to be addressed and overcome before such system is possible. These barriers include the need for a multi-discipline collaborative environment, the collection, integration and analysis of big data and the dissemination of information and tools.

As outlined by the Massachusetts Institute of Technology’s report Convergence: The Future of Health and the European Cooperation in Science and Technology Conference report on personalized medicine, the integration of expertise across multiple disciplines, such as life science, engineering, physics, math and computation science are critical to addressing health care challenges. The report highlights many areas of that will require a collaborative approach to further understanding, such as the need to improve knowledge surround brain function, infectious diseases and cancer. As an example, cancer treatment needs to evolve beyond chemotherapy and radiation. To do so, fundamentally new approaches need to be researched and developed comprising of advanced engineering techniques such as early detection techniques and nano-particle delivery systems as well as genetically guided treatment.

Another barrier to personalized medicine is need for collection, integration and analysis of big data sets. The use of massively parallel sequencing techniques coupled with increasing data collection of clinical, lifestyle and environmental data are all beneficial to the transition to personalized medicine. Regrettably, these data are often inaccessible to large cohort studies, leading to one-off studies that have insufficient power to validate disease risk associations. The lack of consensus, especially in the pathogenicity of genetic variants, has led to confusion and disagreement about the clinical importance of such associations. In an attempt to overcome this barrier, the National Institute of Health initiated the Precision Medicine Initiative, which will collect data from 1 million people across the US. Fostering collaborations with a multidisciplinary approach could allow these data to be continually used for cross-institutional retrospective studies, helping to incentivize updating and achieve a personalized medical system.

Finally, the need for clinician education and access to personalized tools must be addressed. Clinicians and other primary care workers will be instrumental in the transition to a personalized system. However, studies have shown a large gap in needs for a personalized education for front-line workers. In order to implement personalized tools in a clinical setting, it is imperative to help fill this educational gap through easily accessible information and
guidelines. Additionally, clinical decision support tools need to be developed to reduce the knowledge burden on clinicians. Again, this will require a multi-disciplinary and collaborative approach with many stakeholders.

To our knowledge, no system exists to easily facilitate collaboration to foster a precision health care environment, although many barriers exist (Appendix 1).

**PrecMed architecture**

To circumvent many complex privacy and security challenges in the collaborations mentioned above, PrecMed offers a platform to share apps instead of data. The PrecMed apps could be research oriented software, clinical decision supports applications, or patient side support solutions. The PrecMed platform consists of three components: the coordination server, the app store server and the PrecMed client server.

Figure 1 illustrates how the three components are organized and how a typical user would interact. The coordination server hosts the projects and serves as a central platform for coordination among different users during app development and evaluation phases. Compared with traditional research or the software development process, PrecMed offers an open collaboration in a crowdsourcing fashion—each phase can be accomplished by one or more separate projects that are conducted by self-organized groups. After an app is evaluated and approved by a government agency (e.g. FDA), the app can be published to the app store server. The app store server hosts all the approved apps so that any client users can deploy selected apps to their own local PrecMed client server. A PrecMed client server is set up by each healthcare organization or individual practitioner. The PrecMed client server provides a unified environment to host and run locally deployed apps. With user agreement, apps can potentially send aggregated results to the App store server. These results can be used to update mathematic models (e.g. a survival prediction model) and provide feedback on specific populations that may have been originally unavailable in development. Specific components and functions are listed in Table 1. An online interactive mockup can be assessed from https://invis.io/X67R6D6JP

![Figure 1. PrecMed architecture.](image)

<table>
<thead>
<tr>
<th>Table 1. PrecMed components and function descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Components</strong></td>
</tr>
<tr>
<td>Coordination Server</td>
</tr>
<tr>
<td>Component</td>
</tr>
<tr>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Project management component</td>
</tr>
<tr>
<td>IRB coordination component</td>
</tr>
<tr>
<td>Project review component</td>
</tr>
<tr>
<td>Review assessment component</td>
</tr>
<tr>
<td>Credit point component</td>
</tr>
<tr>
<td>Reputation score component</td>
</tr>
<tr>
<td>Approval application component</td>
</tr>
<tr>
<td>App store server</td>
</tr>
<tr>
<td>Web base app store</td>
</tr>
<tr>
<td>Aggregating back end</td>
</tr>
<tr>
<td>PrecMed client server</td>
</tr>
<tr>
<td>App management</td>
</tr>
<tr>
<td>Resource wrapping component</td>
</tr>
<tr>
<td>App execution component</td>
</tr>
</tbody>
</table>

**Innovation and significance**

This framework provides a fundamental change in how personalized medicine information and knowledge is shared by switching from a data sharing model to an application sharing model; potentially bringing a revolution to precision medicine research and practice. The significance is listed as following:

1. Privacy and security issues that have plagued data sharing are addressed by sharing application models and ideas instead of Personal Health Information. PrecMed allows apps to be run in a distributed manner, meaning individual
patient data will be processed without leaving its original site. After the data are processed, aggregated data results can then be returned to PrecMed. This mechanism not only improves the protection of personal health information but also provides application developers real-time feedback to update and improve their app.

2. Using PrecMed, studies can effortlessly reach sufficiently large sample size and be easily conducted across geographic locations. Since data are processed at each PrecMed site, de-identification is not necessary, potentially allowing data access and processing at scale. This is in contrast to many current research studies which, even with a sample size in thousands, is considered to be underpowered. Through collaboration and distributed processing, PrecMed allows for a larger magnitude of data, further pushing healthcare into the big data era.

3. Real-time learning is no longer a dream. Many current precision medicine studies describe prediction models using an array of data resources. Many of these prediction models are trained using a static dataset with few rounds of training. With high volumes of data accumulating every day, it is unreasonable to not utilize the data and update models over time. PrecMed inherently supports learning not only in real-time, but also in distributed manner with the scalability to include hundreds or even thousands of healthcare organizations.

4. PrecMed supports multi-disciplinary team work. Since not every researcher is able to access all the resources that needed to complete their research, PrecMed ask users to contribute what they have and in return offer them what they may be lacking. For instance, an informatician may have expertise in information processing and analyzing, but lacks access to healthcare data. This research can utilize a PrecMed to make an app with sample data and make it available on PrecMed. Other healthcare data holders can review this app and evaluate it on their own site, using their own data.

5. PrecMed offers a centralized location that offers a multitude of collaborative services including: crowd-sourcing idea generation, app development, shared IRB application support, idea and app evaluation, government agency approval and final market dissemination. This service creates a unique place where people can work together toward the goal of precision medicine. Appendix 3 contains multiple use-cases to showcase the PrecMed framework.

6. PrecMed incents users to engage in collaboration through a credit system. As illustrated above in Table 1, this credit system consists of two components: the reputation scores and the credit points.

7. Through reinforcing the PrecMed API (Application Programming Interface) and a set of standards (e.g. FHIR--Fast Healthcare Interoperability Resources standard), PrecMed hides the heterogenous underline IT systems that maintain the health related resources (e.g. personal health record, genetic data, social media data). It offers a unified environment for app development and execution. Thus, PrecMed apps have the outstanding “write once run everywhere” property, which largely maximizes the software production efficiency. As Java virtual machine sits above various operating systems to provide a unified programming environment, PrecMed sits with various health IT systems to provide a unified app developing and execution environment.

8. PrecMed makes app dissemination and deployment easier than ever before. One simple click by an authorized user will deploy the selected app onto the user’s local PrecMed environment. Utilizing user feedback, great apps will receive the attention they deserve and be disseminated quickly, while poorly developed apps will be marginalized.

**Implementation plan**

PrecMed will be implemented through an iterative and incremental development model. In this model, there is an ongoing iterative development circle, which consists of requirement analyses, design, coding, testing and examination. Throughout all phases, a user-centered design principle will be followed in order to encourage participation from all stake-holders including clinical practitioners, clinical researchers, healthcare managers, IRB board members, healthcare government agents, informatics researchers and health IT professionals among others. The architecture of PrecMed will be designed as a platform structure which enables loose coupling between functional components. Development will start with essential components through the incremental modeling process, new and improved functionality will be generated based on continuously gather user requirements and feedback.
Evaluation plan

The evaluation will also be iterative and require multiple phases. For the design phase, evaluation will be conducted through a mockup usability study and/or experts’ walkthrough. In the coding and test phases, evaluation will be conducted through test cases for individual requirements such as those regarding security, validity, reliability, and efficiency. In the examination phase, a test environment will be used. Groups of user’s, including all stakeholders and potential users, will be recruited and asked to use the PrecMed system for a two-month period to evaluate the currently implemented tasks. Users’ satisfaction, engagement, usage barriers, system flaws, and system stability will be studied and evaluated. Hacking competitions will be organized periodically to test the security of PrecMed throughout the iterative development circles, allowing potential security risks to be identified and fixed.

Dissemination plan

Initially, PrecMed will be initiated and developed through a few leading healthcare organizations and published as an open-source platform. Using an open-source platform will allow for more organizations and individual researchers to join during the implementation. Later on, a non-profit organization will be founded to help maintain and disseminate PrecMed. This organization will take charge of PrecMed’s development, evaluation, upgrading, fund rising, project incubating, partner coordination and any other necessary tasks. Books, online tutorials and guidelines will be developed to help new users to get involved. Workshops, seminars, webinars, conferences and hackathons will be organized periodically to increase the PrecMed’s community influence.

Discussion

For precision medicine applications, few alternative solutions are available or feasible. Some public datasets provide structured individual level data (e.g., genotypes and phenotypes), such as dbGaP. These datasets do not include detailed clinical data such as clinical notes. The data structure is fixed and hard to be extended. Some public datasets share structured research results. For instance, ClinVar shares the relationships between human variations and phenotypes. Given current sample size and research scope limitations, many arguable conclusions exist, which overburden the data inclusion decision. The planned NIH initiative for 1 million people data may overcome the disadvantages above; however, concerns are still expressed about data size. Additionally, due to significant costs, data created from previous cohort studies are unlikely to be updated frequently to accommodate new data, procedures, treatments, and tests. Moreover, some data subtlety might be lost during de-identification.

PrecMed overcomes all these weaknesses through sharing applications and ideas, not data. Additionally, it supports real-time data analyses, which is rarely available in the other solutions. One limitation of the PrecMed solution is the possibility of no real data during the app development stage. This shortcoming can be mitigated by using a small amount of de-identified sample data, and following iterative development and evaluation circles.

References

Leveraging High Patient Information Density: Chronic Disease Monitoring with the Precision Patient Ledger

Vincent Caruso, MS, Geoffrey Schau, MS, Kristen Stevens, Matthew Sundling, PhD
Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon

Abstract

Patients undergoing routine monitoring of chronic symptoms are often limited to infrequent interactions with their primary care providers that take place exclusively in the clinical setting. Sparse patient-provider interactions force clinicians to make impactful decisions with equally sparse data. We propose a convenient, easy-to-use digital tool that allows clinicians to prescribe relevant and desirable clinical assessments to be self-reported by patients in the convenience of their own home. By relieving the patient from having to physically visit the clinic to collect meaningful data, we can allow more frequent assessments of patient health. We have developed an example use case of our concept with the Department of Rheumatology at Oregon Health & Science University. The preliminary implementation of our design delivers an assessment of RAPID3, a simple and clinically validated questionnaire approved by the American College of Rheumatology that serves as prognostic assessment for patients living with rheumatoid arthritis. An interactive prototype of our design is accessible at http://precisionpatient.info/.

Introduction

Problem Definition. Presently, patient-provider interactions are typically limited to scheduled clinic visits. As a result, providers are often forced to make major medical decisions based on data collected only during infrequent interactions. Such sparse data may provide an incomplete or even inaccurate description of a patient’s long-term well-being. This issue is compounded for patients who experience difficulty in scheduling, traveling to, or getting time off work to attend their scheduled clinic visit. These challenges implicitly limit the quantity of essential data used by providers to formulate long-term assessments of patient health and develop optimal treatment planning.

A related problem exists for patients who experience chronic yet stable conditions. Inconvenient or difficult circumstances for travel to and from a clinic may yield limited new data to indicate a change in the patient’s condition. The over-treatment of these patients results in inefficiencies by unnecessarily over-scheduling the clinic with visits of marginal value. A tool that allows a patient to regularly conduct a guided self-assessment of one or more chronic conditions and communicate their assessment results remotely to their care provider or electronic health record (EHR) would address both of these related challenges. The opportunity to better personalize treatment plans for chronic disease requires accurate and complete data to describe a patient’s long-term health.

Promoting Home-Based Care through Digitized Self-Assessment. At Oregon Health and Science University (OHSU), the Department of Rheumatology regularly incorporates patient self-assessment a prognostic tool for rheumatoid arthritis patients called RAPID3. The RAPID3 questionnaire asks patients to grade 10 everyday functional activities, such as getting out of bed and lifting a glass to their mouth, on a scale of difficulty, and also asks them to assess pain and global status. Self-reported functional assessments such as RAPID3 have been shown to be the most significant quantitative clinical predictor of premature mortality in rheumatoid arthritis patients when compared to radiographs or laboratory tests

The AMIA Challenge. Our team chose to develop a platform that allows providers to quickly access and interpret long-term trends of their patients’ health, regardless of whether the patient has visited the clinic. We propose a simple, easy-to-use app, which we call the “Precision Patient Ledger,” designed to remotely collect patient-reported health data from clinically-validated health assessment questionnaires including RAPID3. If the long-term trend of a given condition appears stable, then the value of these data could support the decision to decrease the frequency of clinic visits and thus save the patient time and money while reducing the burden on the clinic. Similarly, for patients experiencing a flare-up of their rheumatoid arthritis at the time of their scheduled visit, a provider may access the historical trend leading up to the present event to better determine whether the current case is an acute flare or a chronic indication of worsening prognosis.
Solution Design and Development

*Design Process.* As our AMIA design challenge solution began without clear stakeholders, our initial task centered on identifying a critical need in precision medicine. Each team member proposed a number of clinical problems conducive to data-driven design solutions. In addition to our chosen design solution using electronic self-reporting for chronic diseases management, we also considered (1) an interactive database to allow oncologists and cancer patients to share information on compassionate drug use, (2) an app to assist community oncologists better predict the development of neutropenic fever in at-risk patients, and (3) an app that suggests FDA-approved drugs for repurposing, specifically when patient tumor genomic profiling identifies alterations in the same cancer pathway that an approved drug targets. We ultimately decided to pursue our chosen design solution after conducting a series of stakeholder interviews with clinicians in rheumatology, general internal medicine, medical oncology, and radiation oncology (see Acknowledgements).

This initial exercise laid the foundation for the software design process best suited to the particular strengths of our team members and the constraints of the challenge at hand. Due to the relatively short timeline, we employed many of the principles of Agile Development Methods as described by Ian Sommerville. This approach to software development is based on iterative and concurrent specification, design, implementation, and testing. Importantly, design documentation is minimized to allow greater focus on delivery of a system that is simple and comprehensible on its own (i.e. clear structure, code, and comments). It also utilizes the principle of “people not process,” whereby the skills of the team members are exploited and each person is left to develop their own ways of working in between weekly meetings without prescriptive procedures. Some additional crucial advantages are accelerated delivery of the highest priority functionality to the user with reduced development effort, early user engagement with prototypes to ensure commitment to the system, and focused validation from stakeholders based on the most important system functionality. The prototype of our current design solution is the fourth iteration in this process, reflecting a progression from initial user interface design, static prototype, functional prototype, and finally our user-validated submission.

*Solution Description.* Our proposed solution is a platform that enables patients to regularly self-report data and allows providers to periodically visualize the aggregated longitudinal data. This is comprised of two components. On the patient side, a mobile app permits simple and convenient self-reporting of a prescribed assessment from home or anywhere the patient has access to a smartphone or computer, between visits to the clinic. Only data that has been validated for clinical utility will be collected, maximizing its potential impact. This data is then uploaded to an access-restricted database such as an existing EHR. The second component is an EHR-integrated app that produces a visualization of the longitudinal patient data for the provider, facilitating quick and intuitive review.

The mobile app will fulfill a number of requirements generated in discussions and feedback from clinicians. In our rheumatology prototype, the question and answer format is taken directly from the standard RAPID3 questionnaire app so as to remain faithful to the clinically validated assessment. The app must be simple enough for a wide range of users, especially those who are not computer literate. Our prototype employs a linear design that guides the user from start to finish through a session with the assessment, with easy-to-navigate labeled buttons. Buttons and text on a smartphone are as large as possible to minimize difficulty for patients with rheumatoid arthritis or poor eyesight. Furthermore, while a number of assessments will be possible with the app, to avoid confusion, only those pertinent to a given user will be presented. In order to promote patient self-reporting, the mobile app will include an optional reminder feature, which will prompt patients to complete assessments at appropriate intervals and at a time they deem most convenient. In addition, the app may ultimately include a visualization for the patient of collected data, a potential incentive for those who are motivated by observing their own progress over time. Finally, the app will incorporate its own layer of security, requiring a PIN or login for each use. This will prevent unauthorized data entry by family members who may also have access to the patient’s smartphone or computer.

The other key component for effective use of patient-generated data is a visualization tool for providers (Figure 1). This will be integrated with the EHR as an app, allowing seamless integration into existing clinical workflows. The visualization tool will have an intuitive display, consisting of data points connected by lines plotted against time. This takes advantage of a familiar and easy-to-interpret format for longitudinal data, making patterns and trends immediately apparent to the human eye. The time scale will be adjustable, allowing the user to zoom in or out depending on the desired time frame. In addition, the plot will be minimally annotated with critical patient care events, such as medication changes and missed doses, recent hospitalizations and surgeries. Such annotation serves two important functions: it can alert a provider to potential issues or conflicts with other concurrent treatments; and
it could allow potential inference of the effect of treatments on the patient’s condition based on associated trends in the data.

Almost all clinicians surveyed did not want automatic notifications when patients complete self-assessments, regardless of the resulting score, as this would impose an undue burden on their time and could also create liability concerns. Accordingly, patient data will be readily available to their provider (or a medical assistant) on an “as-requested” basis, typically during the patient visit or just prior. This brings up the importance of educating patients at the time that self-reported data collection is prescribed. Such training could be carried out by a clinic coordinator, nurse or medical assistant who assigns a PIN, demonstrates the app, and advises the patient that their provider is not regularly monitoring their data between appointments. In this way, the app itself will also emphasize patient-provider shared responsibility, suggesting that patients contact their provider directly in the event of a flare-up of their disease.

Alternative Solutions and Comparison

The current standard practice at OHSU and other clinics that have adopted regular RAPID3 assessments is to have patients complete a pencil-and-paper assessment upon arrival at the clinic. This means that patients’ disease status is reported and collected according to appointment frequency, which typically may only be once every few months. While this data is valuable and has been shown to be clinically useful, it leaves significant gaps in the record of disease progression/regression, and in the worst case such sparse reporting may actually misrepresent a patient’s disease. Currently, RAPID3 assessments are recorded in the EHR, but there is no tool for visualization of aggregate data. Instead, scores are tabulated in flowsheets, which give a less intuitive impression of patterns and trends. Furthermore, “between-visit” self-reporting has been shown not only to improve patient satisfaction with their healthcare, but also improve clinical outcomes. One notable case from medical oncology shows a significant difference in 1-year survival between patients using a computerized self-reporting system for symptoms while undergoing chemotherapy treatment compared to usual care.

We found some examples of existing mobile healthcare apps designed to track chronic disease symptoms. One of these is the Orchestra app, developed by the Collaborative Chronic Care Network housed at the Cincinnati Children’s Hospital Medical Center. This app, designed for young patients with irritable bowel syndrome, is similar to our proposal in concept and purpose, but provides tracking of a much wider range of data, including various symptoms, behaviors such as sleep and diet, and journals. It also incorporates messaging and lab chart review features. While such rich data would likely be very useful for personalizing treatment of chronic disease, it also presents drawbacks to widespread adoption. Such a level of self-reporting would only be met by highly motivated patients, while the time required for meaningful review of such diverse personal data would likely overwhelm most already-busy care providers. Furthermore, the complexity of the app would exclude its use by patients with low
computer literacy. Actual use of Orchestra appears quite low, as evidenced by its “100 Downloads” status in the Google Play Store.

Another chronic symptom-tracking tool that launched this Spring as a pilot program is the myHand app by eTreatMD. This app is designed for patients with osteoarthritis. It uses a smartphone camera to measure joint abnormalities and tracks pain and other symptoms. Unlike our proposed platform, it does not claim to track clinically validated metrics, nor does it appear to be intended for integration with the EHR. Rather, myHand appears to be focused on symptom tracking and management by patients themselves, with the possibility of informal sharing of this information with a doctor.

Implementation and Dissemination

The specific design of the Precision Patient Ledger as a secure mobile application focused on three key principles: 1) accessibility, 2) robust communication, and 3) integration with an EHR system. We pursued a design that leveraged existing, and well-established technology components to expedite implementation, and to ensure compatibility with existing EHR platforms and mobile devices. Additionally, we made overarching design decisions to comply with HIPAA requirements in both the mobile health app, and in the communication pipeline of patient data to the EHR.

Accessibility. A critical aspect of success for any mobile healthcare application is accessibility. We focused on creating a patient facing web-based application (or web app), designed to be compatible with a range of mobile browsers, leveraging the pervasiveness of smartphones and tablets. In addition to standard good user interface design practices, we decided on a number of key design elements important to make our app usable by a wide range of people, including patients with limited manual dexterity, color blindness, or poor eyesight. Our goal was to create a simple, intuitive, and consistent application compatible with many mobile devices. The interface was designed to adhere to the original language of the clinically validated tools, to help limit patient confusion and maintain the consistency and validity of the data entry.

We iterated through a series of interface designs, making improvements to the design based on specific feedback on usability, layout of functional elements (e.g. buttons, slider-bar selectors, etc.), color-scheme choice, and adhering to good UI design best practices. The interface design drafts started with 1) a hand-drawn storyboard, converted to 2) a PowerPoint slide show illustrating key application screen shots, then transferred into 3) a dummy interface using an online prototyping service to create a high-fidelity iPhone app mockup, and then implemented as 4) a functional online web-app (http://precisionpatient.info), compatible with most mobile web browsers. Our final prototype was coded using HTML5-compatible mobile-ready HTML, formatted for both desktop and mobile devices using CSS. The interface widgets (buttons, etc.) used the off-the-shelf JavaScript library components found in the jQuery library (v2.2.4). User credentials validation, data validation, and on-device storage was coded using PHP.

Robust Communication. Mobile healthcare applications that communicate over the internet need to be able to compensate for intermittent network access, while maintaining security of patient data, and compliance with HIPAA requirements. Our application required a few basic network functions, including verification of patient identity, secure communication across an insecure network (i.e. Internet), and deposition of mobile app data into EHR records. We decided to compensate for intermittent network access by creating a secure local database to cache application data until network access is reestablished. To meet our other communication requirements, we used the Fast Healthcare Interoperability Resources (FHIR), which is an emerging standard for exchanging medical information electronically. FHIR has a number of interoperability and communication components that are designed to address our application’s secure communication needs. Moreover, it is a software layer that is independent of any particular EHR system, and can be configured to communicate with any common EHR system typically found in a clinical environment. This gives our mobile app the added benefit of general EHR compatibility, and ensures, technologically speaking, that it could be widely used by the medical community.

Integration with an EHR system. A key design decision for the Precision Patient Ledger was to have direct integration of patient data with their electronic health record. This would provide the clinician with on-demand access of the mobile app data through the EHR software interface via an EHR integrated app.

Evaluation

As part of our agile development methods software design process, we have already initiated the first iteration of evaluation of our design solution. We presented our working prototype for faculty review at the OHSU/VA Rheumatology Grand Rounds on June 1, 2016. Here we elicited both oral and written feedback in the form of a
group discussion and short questionnaire. Much of the oral and free-form written feedback has been incorporated into our design solution as described previously. In response to a directed question on usability, Figure 2 demonstrates that 7 out of 10 rheumatologists would use our app (or something similar) in their clinical practice, and 9 out of 10 were open to the idea so long as an education component was included for older patients and those patients with lower computer literacy. Future evaluation will include patient usability feedback targeting the rheumatoid arthritis population.

**Figure 2.** OHSU Rheumatology faculty responses to the question: “Would you use this app (or something similar) in your clinical practice? Please choose Yes/Maybe/No.”

![Graph showing responses](#)

**Conclusion**

Ubiquitous digital devices afford more frequent, self-administered, and clinically impactful assessment of patient health. Incorporating higher-frequency data collection methods without over-burdening either the patient or provider could significantly impact the quality of decisions made within the clinic towards chronic symptom management. Empowering patients with easy-to-use tools for self-monitoring of their own symptoms may lead to a reduction in unnecessary clinic visitation, improve the quality of information needed to accurately monitor patient symptoms, and increase patient involvement with their own data-driven health.

**References**


**Acknowledgements**

Cailin Sibley, MD, Assistant Professor of Medicine, Department of Rheumatology, OHSU
OHSU Rheumatology Faculty
Charles Thomas Jr, MD, PhD, Professor and Chair, Department of Radiation Medicine, OHSU
Erik Fromme, MD, Palliative Medicine, Internal Medicine, OHSU
Joseph Shatzel, MD, Fellow, Hematology and Medical Oncology, OHSU
Steven Kassakian, MD, Internal Medicine, VA Portland, NLM Postdoctoral Fellow, DMICE, OHSU
Eilis Boudreau, MD, PhD, Associate Professor of Medicine, Department of Neurology, DMICE, OHSU
Deborah Woodcock, MBA, Informatics Discovery Lab, DMICE, OHSU
Michelle Hribar, PhD, NLM Postdoctoral Fellow, DMICE, OHSU
Shannon McWeeny, PhD, Associate Professor, DMICE, OHSU
Aaron Cohen, MD, MS, Associate Professor, DMICE, OHSU
Jiri Sklenar, PhD, Research Associate Professor, KCVI, OHSU

1697
Obesity Tracker Application: Mobile Health for Obesity in Children

Nattapon Thanintorn, MS¹, Pericles S. Giannaris, MS², Poungkamon Krisanabud, MS³
¹Department of Pathology and Anatomical Sciences, ²Informatics Institute, ³Sinclair School of Nursing
University of Missouri, Columbia, MO

Background
In the United States, the estimated one third of or 12.7 million children and adolescents are obese.² In addition to genetics, behaviors, and environments, adult of household’s education level and low-level income are associated with development of childhood obesity.³ Children are vulnerable and lack an understanding for their healthcare and food consumption provided by parents. It is common that children are diagnosed with obesity at later stage of life. Obese children have increased risks to develop comorbidities, such as cardiovascular disease, hypertension, prediabetes, and musculoskeletal problems.⁴, ⁵, ⁶ Obese children are more likely to be obese adults with lifelong comorbidities. For obesity, it is difficult for clinicians to diagnose if newborns will become overweight or obese due to personalized molecular mechanisms as well as other uncontrollable risk factors.

To reduce and prevent childhood obesity development and comorbidities, integration of mobile technology, advanced technology in clinical systems, and evidence-based guidelines has been introduced to shed some light on better tracking for children growth and more effective communications between healthcare providers and patients. The use of health mobile applications on smartphone contribute to increased physical activity for adults.⁷ Utilizing mobile technology to promote healthcare for obese children and family is promising because cell phones or tablets become one of the necessities in their lives. According to Pew Research Center, in 2013 estimated 37% of children aged between 12 to 17 years have a smartphone. Mobile health application development for children and adolescents to promote interaction between clinicians and patients, healthy behaviors and awareness of their health may contribute to substantially decreased childhood obesity.

However, WHO Anthro, one of the existing tools, is a good example of why we need to provide a solution for the AMIA specific challenge 2016. The tool is designed for healthcare specialists and lacks a communication space for interaction between healthcare providers and patients. The results and visualizations are not intuitive and easily understood by children or adolescents.

Proposed Solutions
In response to the AMIA specific challenge, we propose a mobile application, Obesity Track Application (OTA), that provides i) seamless interaction between providers and children with tendency of obesity development, ii) prediction of obesity development at birth using evidence-based guidelines and electronic health records (EHRs), and iii) visualizations on child’s growth and obesity development progress. The rationale for the proposal is that once children or adolescents understand risk factors and patient-specific explanatory growth chart that narratively depicts level of obesity and comorbidities, resulting in innovative approaches to promote healthier behaviors and prevent obesity development.

For providers, this application will allow them to log in on mobile devices to keep track on patient’s obesity development, receive automatic alerts when body-mass-index z-score substantially exceeds the standard BMI, and provide recommendations to patients. For patients, this application will allow them to engage interaction with clinicians through contact module and visualize their growth chart. In short, this application can facilitate shared patient-provider shared decision making in precision medicine setting and enable monitoring of growth chart and obesity development overtime using mobile technology and EHRs, and provide more self-explanatory displays for child patients.
Design Process

The application requires a username and a password for healthcare providers and patients. We design two interfaces for types of users: 1) patients (Figure 1) and 2) healthcare providers (Figure 2).

![Figure 1. Interface for patient.](image1)

![Figure 2. Interface for doctor.](image2)

Records

Patients have records to store patient’s weight and height over time. A record is evaluated as underweight, overweight, obese, healthy states based on BMI z-score that depends on age (taken from WHO standard).

Growth Chart

Growth chart is generated from the values of patient records and bmi-for-age z-scores taken from the WHO website.
Blood Tests
Blood test menu is designed to display recent test results.

Recommendations
Recommendation from doctors is a part of doctor-patient interaction. A doctor should give a recommendation to a patient after evaluating patient’s information if necessary. Getting a recommendation from doctor encourages parents and patients to continue updating the growth record and aware that their doctor is regularly monitoring the patient’s progress.

Contact Requests
Contact requests menu is important for doctor-patient interaction. A patient/parent can send a contact request by choosing a topic when contact is needed, thus doctors can be notified and contact during available time. A patient/parent can send only one contact request till doctors mark that request as contacted. This prevents doctor from getting too many requests from one patient.

Obesity Risk At Birth
Calculating obesity risk at birth is an approach to predict child obesity at birth. This probability is calculated by Parents BMI, Number of household members, Mother’s profession category, gestational smoking and weight at birth.

For the outcomes, OTA is expected to provide a virtual space for healthcare providers and patients with evidence-based guidelines and self-explanatory visualization to promote an important positive impact on optimal theranostic strategies in precision medicine setting.

Implementation and Dissemination
The first prototype of OTA has been implemented and completed using .NET framework within Laboratory of Translational Bioinformatics at University of Missouri. This pilot program will be tested with pediatricians and selected children with obesity condition or their family. The program requires a username and a password to access and establish communications between healthcare providers and patients. Visualizations and interface designs will be improved based on feedbacks users from the pilot testing. For a large-scale dissemination, the application will be available for download for android and apple devices. Online tutorial for the application will be presented to the medical community.

Solutions Comparisons
The OTA mobile application is useful to keep records of growth and generate growth chart for children and adolescents from 0 to 19 years of age. WHO Anthro is specifically designed for healthcare specialists. The results require time to comprehend while the OTA application provides more self-explanatory visualizations for patients

Table 1. Strengths and Weaknesses for WHO Anthro Program.

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Detailed calculations</td>
<td>• Difficult to learn</td>
</tr>
<tr>
<td>• Detail and variety of charts</td>
<td>• Efficient only for 0-60 month-old children</td>
</tr>
<tr>
<td>• Detailed child visit records</td>
<td>• No doctor – patient interaction</td>
</tr>
<tr>
<td>• Creating reports</td>
<td></td>
</tr>
<tr>
<td>• Useful for doctors to keep track of child till 60 months of age</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4. Visualization of BMI analysis from Obesity Tracking Application.

Figure 5. Output from WHO Anthro

Figure 6. Visualization of Growth Chart from Obesity Tracking Application.

Figure 7. Visualization of Growth Chart from WHO Anthro.
Conclusion

This application can be used as an effective tool to motivate parents to accept child’s situation and decide to make changes on child’s lifestyle. If their parents use this mobile application and monitor child’s growth and situation statistically, they can be more confident to make changes for their children. This application also makes parents and children feel taken care of because their doctors regularly monitor their progress and situation.

References

SocialVue: Making Social Determinants of Health Visible in the EHR

Steven R. Chamberlin, ND, MAC, Ian T. Stavros, BS, Dana Womack, MS, RN
Oregon Health and Science University, Portland, Oregon

Introduction

Every patient has a distinctive physiological make up, and every patient has a unique life story. In precision medicine, clinicians strive to tailor care to unique patient characteristics. While the term precision medicine (PM) is often used synonymously with genetics-informed treatment, social determinants of health (SDoH) are social and behavioral characteristics that also affect a patient’s quality and longevity of life. Inclusion of SDoH in the electronic health record (EHR) can provide vital information to inform and improve clinical care for individuals, public health of populations, and SDoH research(1).

Although there is a growing body of evidence regarding the effects of social and behavioral characteristics on the onset and progression of disease, clinical care is not yet routinely targeted to patients’ individual SDoH characteristics. Contributing factors include a lack of SDoH documentation in the electronic health record and the absence of tools to support meaningful discussions about SDoH between providers and patients. In this student design challenge, we propose creation of an EHR-compatible application that provides at-a-glance awareness of patient SDoH issues and facilitates delivery of care that is targeted to a patient’s social and behavioral circumstances. A prototype of the application can be accessed at http://tinyurl.com/SDoH-2016.

The World Health Organization defines SDoH as the “conditions in which people are born, grow, live, work, and age, and the wider set of forces and systems shaping the conditions of daily life (2)”. In 2014, the Institute of Medicine (IOM) recommended eleven social and behavioral domains and twelve associated measures for inclusion in EHRs, but did not specify how data should be captured, presented, or incorporated into tools to support patient-provider engagement (1). The IOM report focuses primarily on patient reported SDoH data, but the literature also suggests inclusion of geography based publicly available data in the EHR, often referred to as “community vital signs” (3, 4). Informatics issues include data capture, standardization, storage, privacy, interpretation and reuse(5). There is currently a paucity of literature regarding how to present SDoH data within the EHR, and how to make this actionable at the point of care.

We iteratively collaborated with clinicians to addresses the following opportunities:

1) Integration of clinical, community, and patient-specific SDoH data in the context of the EHR
2) Increased clinician awareness of patient’s unique social and behavioral characteristics
3) Improved engagement between clinicians and patients re: unique SDoH characteristics in a manner that leads to targeted care

Design Process

We began our design process with the goal of facilitating the delivery of precision medicine by tailoring care to patients’ SDoH characteristics, but we did not have a preconceived notion of what the solution might entail. We relied heavily on representative end users to inform our resulting design solution.

Problem addressed. During an early design session at a federally-qualified health center, a social worker noted that making the SDoH data she routinely collects from patients visible to providers and other team members in the medical home is a pressing challenge. She shared that she has tried multiple locations in the EHR to help clinicians “start seeing” SDoH, including a care coordination note at the top of the problem list, a dedicated clinical note, patient history form, and the “barriers” section of the chart. She reported that none of these locations has been successful and suggested that a novel approach for making clinicians aware of unique SDoH patient characteristics is needed. We selected this as our target problem as this challenge was echoed by multiple clinicians we interviewed.

Target end users: Our target end users are clinical members of multidisciplinary teams in the medical home setting (e.g. providers, social workers, nurses, medical assistants, others). We selected this user group because multidisciplinary teams are well suited to address social determinants and deliver precision medicine in the context of the medical home.
**Design objective.** The objective of our design is to provide medical home team members with at-a-glance awareness of patients’ SDoH characteristics, facilitate engagement between providers and patients regarding SDoH characteristics, and support care that is targeted to a patient’s social, economic and environmental circumstances.

**Target clinical condition:** Our resulting solution could be applicable to multiple clinical conditions, but we selected Type II Diabetes as our target clinical condition. Our rationale is that diabetes is highly influenced by patient’s social and behavioral characteristics. Limiting the focus to a single condition scopes this project to a manageable size, and demonstrating feasibility and value in the context of a single condition would be helpful before expanding the design concept to support multiple clinical conditions.

**Design approach.** We employed Agile & User Centered Design Integration (AUCDI) as our design approach, as AUCDI incorporates the principles of user-centered design within the context of an agile development process (6-8). In user-centered design, representative end users actively participate in the design process to guide the development of functional requirements and to participate in design and usability testing of a product through all states of development. Agile development is an iterative development process in which a product is built across multiple time-boxed windows, called sprints, with the goal of developing a minimally viable product as quickly as possible. We selected AUCDI as a method that blends user centered design and agile development because it supports evolutionary requirements elicitation and incremental design via a short and iterative feedback loop between the design team and target end users.

**Requirements elicitation, prototype design, and initial evaluation.** Recognizing that effective design begins with a clearly understood problem and active engagement with end users, we executed multiple interactive design sessions with representative stakeholders across three 2-week agile design sprints, following our selected AUCDI design methodology. We conducted design sessions with at least 5 representative end users during each sprint, for a total of 16 interactive sessions with these users. We are grateful for substantial end user input, and have included the names, organizations and roles of design participants in Supplementary Materials Exhibit 2.

- **Sprint 1:** In our first design sprint, we developed a low-fidelity prototype as a launching point for design sessions as it can be easier for users to react to a loose idea than to “blue sky” new capabilities. Our interactive discussions focused on clarifying our design goals and identifying what SDoH information is most supportive of SDoH-targeted care. Key learnings from this sprint included the need to prioritize collection and display of patient-specific SDoH data (e.g. patient experiencing food insecurity), over community-level data (e.g. percent of neighborhood under 200% poverty level). Representative end users found community data to be interesting, but it is patient-specific SDoH characteristics that they value most. Users reported that patient-specific SDoH data is typically collected by a social workers or medical assistant for subsequent use by all team members in a medical home. Users emphasized the need to make SDoH highlights available to providers at a glance in the EHR, allowing them to enter a patient room with awareness of a patient’s SDoH characteristics.

- **Sprint 2:** In sprint two, we developed a higher-fidelity prototype that included the concept of a “hover” icon that hovers of selected EHR screens and when moused over, expands to provide immediate access to patient-specific SDoH highlights. We also introduced an interactive map of a 1-mile radius around a patient’s residence that highlights community features of interest to diabetes including parks, grocery stores, fast food restaurants, and pharmacies. During interactive design sessions, users validated that the hover icon was helpful and provided input on what it should look like, and the content it should display. Users appreciated that SDoH highlights were not hidden behind “yet another tab” in the EHR, and provided considerable input regarding how to organize SDoH under relevant headers. Users also identified a small subset of publicly available community vital signs that would be useful when tailoring care to patient’s SDoH characteristics. We revised our evolving prototype frequently throughout the entire process, but in this sprint in particular, it was revised on nearly a daily basis.

- **Sprint 3:** Our third design sprint focused on preliminary validation of the prototype design and elicitation of feedback regarding usability. By this time, we had migrated from a low-fidelity prototype to a functioning web-based prototype that users could interact with. Users provided feedback on the functionality of the prototype and the proposed workflow which resulted in additional refinements, such as the ability to expand the map and trended clinical data to a “Patient View” mode. This could facilitate selection of a pharmacy that is on a patient’s bus route or open during a time of day that the patient is off work, setting exercise goals that incorporate a specific park or other landmark, or discussion of the relative benefit of different food choices that exist in a patient’s immediate neighborhood. By the end of this sprint, end users became increasingly enthusiastic about the design and started asking the team when it may become available for use in practice.
Discussion of Alternative Solutions Considered

This design team considered multiple design alternatives (Table 1), including designing additional functionality within the framework of an existing EHR, development of a stand-alone app, and an EHR-compatible app that can work in conjunction with any EHR as overlay functionality. Of the alternatives considered, the EHR overlay app is the most generalizable solution in that it could operate across multiple EHR systems.

Table 1. Alternative Solutions Considered.

<table>
<thead>
<tr>
<th>Alternative Considered</th>
<th>Strengths</th>
<th>Weaknesses</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design an additional data capture screen in an existing EHR</td>
<td>Could be made available to clinicians in a relatively short time frame</td>
<td>Limited to a single EHR solution; not reusable across vendors. Design is limited to existing EHR functionality</td>
<td>We rejected this option because users value a community map, which exceeds the capabilities of EHRs today.</td>
</tr>
<tr>
<td>Design a stand-alone app to display community vital signs</td>
<td>Standalone applications are faster to develop than applications integrated with an EHR.</td>
<td>Necessitates a separate workflow as a standalone system. Lacks integration with clinical EHR data.</td>
<td>We rejected this option because users value patient-specific SDoH data over community vital signs data.</td>
</tr>
<tr>
<td>Design an EHR-compatible app that provides novel visualization, patient-specific SDoH data, selected community data, and clinical data</td>
<td>Tells a more complete patient story, integrates clinical, pt-specific and community-level SDoH data. Provides maximum flexibility and widest distribution across EHRs.</td>
<td>Longer development time than building additional screens in an EHR. Clinic IT departments will need to approve and maintain use of an EHR-compatible app.</td>
<td>We selected this option because it includes integration of patient-specific, community-level and clinical data, and because an EHR-compatible app can be distributed across multiple EHR systems.</td>
</tr>
</tbody>
</table>

Proposed Solution

The proposed solution, SocialVue, is an EHR overlay app that will help providers quickly digest SDoH data and tailor care to patients’ unique characteristics. We describe key functionality through the lens of a patient story, using italics to highlight examples of enhanced engagement and SDOH-targeted care.

Patient story: Natalia is a 56 year old Latina female with type II diabetes. She arrives for a same-day appointment, because she is concerned about a scratch on her leg that may be infected. Her primary care provider is out of the office, so she is seen by Dr. Smith. Not familiar with Natalia’s case, Dr. Smith opens her chart to review her medical history upon entering the patient room. During a previous visit, Natalia spoke with the medical home’s social worker who utilized SocialVue to note Natalia’s specific SDoH characteristics.

- **SDoH hover icon.** When Dr. Smith opens Natalia’s record, he immediately notices an SDoH icon hovering over the home screen, alerting him that Natalia has SDoH characteristics that may affect her clinical care (Figure 1). He knows that this icon only appears when known issues have been identified, typically during past patient visits with the medical home’s social worker. As a result, Dr. Smith has come to trust that the icon’s presence means that useful information is available.

  ![Figure 1. SDoH hover icon, generated using Piktochart software, and displayed over a generic EHR screen, publicly available at http://technologyadvice.com/blog/healthcare/3-hospital-ehr-solutions/](http://technologyadvice.com/blog/healthcare/3-hospital-ehr-solutions/)

- **SDoH highlights at-a-glance.** When Dr. Smith hovers over the icon, it expands to provide immediate access Natalia’s SDoH highlights (Figure 2), populated during a previous visit via an SDoH intake form, shown in Supplemental Material Exhibit 4. He notes that Natalia is looking for work, has food insecurity, and has recently moved in with a friend. In the time it takes to hover over an icon, Dr. Smith has gone from knowing nothing about Natalia, to understanding her key SDoH challenges and her personal goal related to managing her diabetes. This provides a framework for the visit and allows Dr. Smith to tailor care and incorporate SDoH information into the clinical encounter.

  ![Figure 2. Expanded SDoH hover icon](http://technologyadvice.com/blog/healthcare/3-hospital-ehr-solutions/)

1705
Natalia explains that she was scratched by some thorny bushes when she visited an unfamiliar park. Dr. Smith acknowledges that he understands she moved recently, and commends her on maintaining her goal of walking around a park three times/week at her new location, as he begins to examine her leg. He determines that her leg is indeed infected and recognizes that he will need to order an antibiotic.

- **Diabetes 360 screen**, with integrated clinical, SDoH, and community data. Dr. Smith also wants to check on Natalia’s diabetes, so he clicks the SDoH icon to bring up a Diabetes 360 screen that displays an integrated view of Natalia’s clinical context, more detailed SDoH information, and her community context (Figure 3). Visual cues alert him to the fact that her blood pressure is elevated, so he emphasizes the value of continued exercise to manage both her blood pressure and diabetes. Together, they look at the map to **identify additional parks in her neighborhood that Natalia could explore also identify healthy and affordable food sources in her neighborhood**.

Having awareness of Natalia’s unique SDoH characteristics, Dr. Smith recognizes that it will be important to patient compliance to select a pharmacy near Natalia’s new residence, so he clicks “Patient View” to enlarge the map that covers a 1-mile radius around her address (Supplementary Materials Exhibit 1). Together Dr. Smith and Natalia **identify a pharmacy in her neighborhood that is convenient to public transportation**. By hovering over the pharmacy icon, they are able to identify that the pharmacy is ¼ mile from Natalia’s residence, and that it is open until 9 pm daily.

![Figure 3. Diabetes 360 screen, incorporating clinical, patient-specific SDoH data, and community-level vital signs](image)

- **SDoH input screen**. Natalia’s SDoH information on the Diabetes 360 screen (Figure 3) is populated via an SDoH intake form (Supplementary Materials Exhibit 2). The envisioned workflow, developed in collaboration with our clinical end users, is that the intake form would be initially filled out, and updated by medical assistants during the rooming procedure, or during patient interactions with a social worker.

- **Integration with the EHR**. While the data in our current prototype is simulated data, the envisioned solution will exist as an overlay app that is integrated with an EHR system, such that the clinical data on the Diabetes 360 screen is populated directly from the EHR. Next steps include a due diligence investigation to assess to what extent the SMART Health IT platform and Fast Healthcare Interoperability Resource (FHIR) data models(9) support the data elements required by the proposed solution. Additional future work includes formal usability testing of the interactive prototype, development into a fully developed application, and inclusion of additional community resources (e.g. food pantries) on the map.
Implementation, Dissemination and Evaluation Plan

The design team has discussed implementation and dissemination of SocialVue with OCHIN(10), a nonprofit organization in Oregon that provides EHR functionality to over 500 clinics that serve underserved patient populations in 18 US states. The SocialVue app is synergistic with OCHIN’s desire to incorporate SDoH data into its EHR in ways that support targeted care. OCHIN has expressed a sincere interest in taking prototype developed through this project, and developing it into a deployable app that sits on top of the EHR that OCHIN provides to its clinic clients. OCHIN may include SocialVue in a multi-arm study designed to evaluate the relative effectiveness of different ways of presenting SDoH data to end users in the EHR and evaluate impact on clinical outcomes.

Prior to deployment to selected OCHIN clinics, SocialVue will undergo rigorous usability testing, beyond the interactive sessions executed by the student design team during the initial development of the design concept. During usability testing, representative clinical end users will be observed as they attempt to execute defined tasks using the app. Users will be encouraged to talk aloud as they execute these tasks, so that designers can understand both what users are thinking, and what aspects of the design or user interface are causing challenges. Usability testing will lead to system enhancements that will be made prior to rollout to actual end users.

Clinical outcome evaluation related to use of the SocialVue app will be assessed as part of a planned future OCHIN study. A selected number of clinics will be randomized to use either SocialVue or one of two additional competing designs, to be developed by OCHIN. OCHIN plans to assess clinician satisfaction regarding use of SDoH tools through a survey and clinical outcomes will be compared across sites. If SocialVue is shown to improve targeted care and clinical outcomes, there is potential for it to be deployed to all clinics that utilize the OCHIN EHR.

Summary

In this submission, we present a novel approach for targeting care to patient- community-level SDoH data in the context of the EHR. Use of a combined agile & user centered design approach allowed representative users to provide substantial input across 16 interactive design sessions. Our functioning prototype demonstrates how an EHR overlay app can call attention to SDoH characteristics via an icon that expands to provide immediate access to SDoH highlights. Together, the icon and interactive “Diabetes 360 screen” create an SDoH framework for the visit in the context of an existing EHR workflow that facilitates patient engagement and results in SDoH-targeted care.

References

Content and Quality of Free-Text Occupation Documentation in the Electronic Health Record

Ranyah Aldekhyyel, MS1,5, Elizabeth S. Chen, PhD2, Sripriya Rajamani, MBBS, PhD, MPH1,3, Yan Wang, PhD1, Genevieve B. Melton, MD, PhD1,4

1Institute for Health Informatics, 3Public Health Informatics Program, and 4Department of Surgery, University of Minnesota, Minneapolis, MN; 5Medical Education Department, College of Medicine, King Saud University, Riyadh, SA

Abstract

Recent recommendations for capturing social and behavioral information in electronic health record (EHR) systems for downstream applications, including research, highlight the need to better represent patient occupation. The objectives of this study were to characterize the content and quality of EHR social history module free-text occupation documentation. After developing categorization schemas, occupation entries with frequencies >5 (n=2,336) and a random sample of those with frequencies ≤5 (n=381) were analyzed. The information contained in the 2,336 entries fell into five groups: occupation (84.7%), occupation details (20.6%), employment status (2.5%), not in labor force (21.6%), and other (2.5%). Quality issues included use of acronyms/abbreviations (9.1%) and misspellings (1.6%). In comparison, quality issues with the 381 entries were: other (29.1%), acronyms/abbreviations (19.0%), and misspellings (9.0%). These findings suggest the need for EHR user training, system enhancements, and content standardization to support use of occupational information for clinical care and research.

Introduction

With the increasing use of Electronic Health Record (EHR) systems driven by various healthcare reform initiatives and the EHR Incentive Program1, there is an opportunity for enhanced capture of occupation data electronically at the point of care. The importance of documenting social and behavioral factors influencing health status and outcomes has been recognized and supported by recommendations from respected advisory bodies and organizations, such as the National Academy of Medicine (NAM; formerly Institute of Medicine) and the National Institute for Occupational Safety and Health (NIOSH). For instance, in 2011 the NAM published a report entitled “Incorporating Occupational Information in Electronic Health Records,”2 which highlighted the need for representing occupational information in EHRs through emphasizing the potential benefits to the individual patient as well as the population as a whole, and included recommendations for next steps. The report also illustrated several examples suggesting that the presence of the patient’s type of work in the EHR could enable more accurate diagnosis and treatment of specific medical problems, which could lead to improved quality and efficiency of care. NIOSH supported incorporating occupation in the EHR through publishing demonstration projects, which focused on the representation of occupation related information. These include the Occupational Data for Health (ODH) data model3 (work in progress, currently not available in the public domain), HL7 Clinical Document Architecture (CDA) standard template for “Occupation Data for Health”4, and various pilot projects to understand and promote the capture of occupation data5. All these resources point to occupation as a complex concept with inter-related elements (e.g., occupation, industry, employer, and employment status).

In 2014, follow-up recommendations for capturing social and behavioral information in EHRs were issued in two reports published by the NAM5,6. Occupation/Employment was considered under the socio-demographics domain in the first report (“Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1”) and employment was identified as a candidate domain in the second report (“Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2”). The report also indicated an instrument to standardize the collection of employment information, which was the Multi-Ethnic Study of Atherosclerosis (MESA) question and categories5 for employment status.

Prior work has focused on reviewing the capture of social and behavioral factors in the EHR, involved characterizing social history information in clinical notes and identifying the eight most common statement types, one of which was occupation which represented 10-15% of statements across three different sources of clinical notes7. Subsequently, social and behavioral information was also examined in three public health surveys: Behavioral Risk Factor Surveillance System (BRFSS), National Health and Nutrition Examination Survey (NHANES), and National Health Interview Survey (NHIS)8. Occupation was a common survey item with 22 questions and corresponding responses across the three surveys.
Building upon findings from prior work and national recommendations for capture of occupational information in the EHR, the main objective of this research is to understand the practice of documentation of occupation in an institutional EHR by analyzing contents of free-text occupation entries assessing the type of information captured and evaluating the quality of data stored. This study uses similar methodologies for analysis as developed and applied in previous studies that focused on representation of various aspects of social history documentation in the EHR, including tobacco use, alcohol use, drug use and living conditions. Specifically, the aim of this study is to determine issues associated with current representation of occupation in an EHR system with potential implications for enhancing system design for discrete data collection, user training, and standardization of data as well as informing the development of natural language processing (NLP) tools to enhance access to and use of structured data for research, clinical care, and population health purposes.

Methods

Study Design

This study involved a retrospective analysis of data collected in the free-text occupation field within the Epic EHR implemented at University of Minnesota (UMN)-affiliated Fairview Health Services accessible through the UMN Clinical Data Repository (CDR). Occupation and related information is collected in the socio-economic section in the social history module of the EHR and is comprised of one structured field (Employer) and two free-text fields (Occupation and Comments) (Table 1). The current study focused on the “occupation” field and was conducted in two phases for: (1) data extraction, preparation and transformation (Figure 1A) and (2) categorization schema development and application (Figure 1B).

Table 1. Example Occupation Entries.

<table>
<thead>
<tr>
<th>Field</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation</td>
<td>Truck Driver</td>
<td>asphalt / roofing, snow plowing in winter</td>
<td>10th grader - PLSHS</td>
</tr>
<tr>
<td>Employer</td>
<td>Wal-Mart</td>
<td>Other</td>
<td>NONE</td>
</tr>
<tr>
<td>Comments</td>
<td>Night Shift</td>
<td>Increased physical activity in summers</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 1: Study Methodology with Two Phases for Extracting, Preparing and Transforming Data (A) and Developing and Applying Categorization Schema (B).
Phase 1: Data Extraction, Preparation and Transformation (Figure 1A)

Occupation data from the UMN CDR collected from May 2000 through February 2015 were extracted resulting in 3,572,036 entries and imported into a relational database (MySQL). Removal of duplicate patient encounters (n=3,271,909) resulted overall in 300,127 entries. Many empty entries were found in the occupation field, which were then removed along with “0” values, as these entries were considered missing data. All entries were then converted to lowercase in order to select the “unique” entries stored in the free-text occupation field. A “unique” entry is an entry with no exact match within the dataset, taking into account space between words and punctuation marks (e.g., “homemaker” is considered a different entry than “home-maker”). The resulting dataset included 53,424 unique occupation entries representing 185,891 overall entries. The data were then divided based on the frequencies of unique occupation entries. A frequency of “5” was chosen as the cutoff for the two datasets representing about 90% of overall entries (25% of unique entries) excluding those with a frequency of “1”. The first dataset “freq. > 5” was comprised of unique entries with frequencies of more than 5 (n=2,336) representing about 87% of the overall dataset (n=123,789). A random sample of the second dataset was extracted from the remaining dataset (5 or less) of 51,088 total unique occupation entries. The second dataset “freq. ≤ 5” included 381 unique occupation entries and was based on the total number of unique occupation entries with frequencies of 5 or less to provide a precision of 5% at the 95% confidence level.

Phase 2: Categorization Schema Development and Application (Figure 1B)

Schema Development

Categorization schemas were iteratively developed to manually categorize the “contents” of the unique occupation entries and identify any “data quality” issues with the entries. The “contents” schema was developed for identifying the unique contents of the free-text occupation entries. The “data quality” schema was developed to determine any data quality issues with the “contents” in the free-text occupation entries. Both of the developed schemas allowed for entries to be placed under several categories. The process of developing these schemas was based on an approach used in a prior study focused on analyzing free-text tobacco use documentation in the EHR. The general method for developing these guidelines consisted of two phases. The first phase focused on developing initial categorization schemas based on analysis of 50 unique occupation entries and enhancing the schemas through weekly meetings and discussions involving four subject matter experts in the field of informatics with experience and expertise in clinical care, public health, and standards (GMM, ESC, RA, and SR). The second phase involved calculating inter-rater reliability using the kappa statistic to ensure consistency in categorizing entries between two reviewers (RA and SR) using the final versions of each categorization schema for 250 unique occupation entries from the “freq. > 5” dataset (n=2,336).

Earlier versions of the “contents” schema consisted of three main groups with twelve different categories. A notable topic of discussion among the group was focused on identifying the most applicable category for student-related entries. After review of six different sources of information (Standard Occupational Classification [SOC] System from the Bureau of Labor Statistics in the United States (U.S.) Department of Labor, Systematized Nomenclature of Medicine-Clinical Terms [SNOMED-CT], North American Industry Classification System (NAICS) from the U.S. Census Bureau, MESA, NIOSH and MetaMap) and then categorizing 50 entries from the “freq. > 5” dataset, the main groups “not in labor force” and “employment status” were created with associated categories and subcategories. The final version included five main groups,十二 categories, and four subcategories (Table 2). Categories and subcategories were created to provide further detail about the content being documented in the occupation entries. Subcategories are directly linked with a specific category (e.g., the subcategory “Type of Occupation” is directly linked to the “Name of Occupation” category and the subcategories “Student Related –Status”, “Student Related –Type”, and “Student Related –Other” are directly linked to the “Not in Labor Force –Student” category). This means that no entries can be placed under a specific subcategory without having part of the entry categorized under the related category. An example that explains the relationship between categories and subcategories is “Travel Consultant” where the entry was divided into two words: “Consultant” was categorized as “Name of Occupation” and “Travel” was categorized as “Type of Occupation”. Another example is “Law Student” where “Student” was categorized as “Not in Labor Force-Student” and “Law” was categorized as “Student Related-Type”.

The “data quality” schema underwent less revision since the categories were more distinct and captured common issues that are found in free-text data. The final version consisted of five main issues: (1) Misspelling, (2) Acronym/Abbreviation, (3) Ambiguous, (4) Multiple Terms, and (5) Other (Table 3). Two reviewers analyzed a set of 250 entries using the final versions of the categorization schemas. Inter-rater reliability was calculated using Cohen’s Kappa, achieving κ of 0.94 for contents and 0.86 for data quality issues (percentage agreement of 0.99 and 0.98 respectively). All differences in categorization, between the two reviewers, were revised and resolved prior to applying the schema on the datasets.
### Table 2: Categorization Schema for “Contents”

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>Brief Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Group (1): Occupation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Name of Occupation</td>
<td>Describes what kind of work the patient does</td>
<td>• Teacher</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Technician</td>
</tr>
<tr>
<td>1.1</td>
<td>Type of Occupation</td>
<td>Describes the type of a specific occupation. This code is directly linked to name of occupation.</td>
<td>• Electrical Technician - code as (1 and 1.1)</td>
</tr>
<tr>
<td></td>
<td><strong>Group (2): Occupation Details</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Industry</td>
<td>Describes the type of work the patient’s employer or business does. The large perspective of the work sector.</td>
<td>• Human Resources</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Customer Service</td>
</tr>
<tr>
<td>3</td>
<td>Workplace</td>
<td>Describes the place or location where the patient works.</td>
<td>• Daycare</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Warehouse</td>
</tr>
<tr>
<td>4</td>
<td>Job Duties</td>
<td>Describes the activity that the patient is performing as part of an occupation/job. Detailed specific task.</td>
<td>• Office work</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Data Entry</td>
</tr>
<tr>
<td>5</td>
<td>Employer Name</td>
<td>The name of the patient’s employer</td>
<td>• Target</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Fairview</td>
</tr>
<tr>
<td>6</td>
<td>Equipment</td>
<td>Describes the necessary equipment for a particular occupation</td>
<td>• Computer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Heavy equipment</td>
</tr>
<tr>
<td></td>
<td><strong>Group (3): Employment Status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Employment Status</td>
<td>Describes that patient’s current employment status</td>
<td>• Volunteer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Part time</td>
</tr>
<tr>
<td>8</td>
<td>Unemployment Status</td>
<td>Describes that patient’s unemployment status</td>
<td>• Unemployed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Currently unemployed</td>
</tr>
<tr>
<td></td>
<td><strong>Group (4): Not in Labor Force</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Not in Labor Force –Student</td>
<td>“Students” who are neither employed nor unemployed</td>
<td>• Student</td>
</tr>
<tr>
<td>9.1</td>
<td>Student Related –Status</td>
<td>Describes the student’s current enrollment status. This code is directly linked with “student”.</td>
<td>• Full time Student - code as (9 and 9.1)</td>
</tr>
<tr>
<td>9.2</td>
<td>Student Related –Type</td>
<td>Describes the student’s major at school. This code is directly linked with “student”.</td>
<td>• Nursing Student - code as (9 and 9.2)</td>
</tr>
<tr>
<td>9.3</td>
<td>Student Related –Other</td>
<td>Describes the grade or level of student. Does not have to be associated with student.</td>
<td>• 1st grade</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Sophomore</td>
</tr>
<tr>
<td>10</td>
<td>Not in Labor Force –Other</td>
<td>Persons who are neither employed nor unemployed and are not students</td>
<td>• Retired</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Home maker</td>
</tr>
<tr>
<td></td>
<td><strong>Group (5): Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>NA/None</td>
<td>Includes “NA” and “None”</td>
<td>• NA, None</td>
</tr>
<tr>
<td>12</td>
<td>Miscellaneous</td>
<td>Includes numeric values, type or status of patient, “other”. Provide details as part of comment/notes.</td>
<td>• Child, Kid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Single</td>
</tr>
</tbody>
</table>

**Schema Application**

Coders manually categorized unique occupation entries using the two developed categorization schemas. The remaining 2,036 set of occupation entries, from the “freq. > 5” dataset (n=2,336), in addition to the “freq. ≤ 5” dataset (n=381) were divided in half between the same two reviewers who previously categorized the 250 unique occupations, to apply the developed schemas.

“PHY ED TEACHER” is an example of an entry from the “freq. > 5” dataset. This entry was categorized as “name of occupation” and “type of occupation” for “contents” and the “data quality” issues identified were “acronym/abbreviation” and “other-uppercase letters”. Other examples of entries include “unemp” that was categorized as “unemployment status” with one identified quality issue of “acronym/abbreviation” and “Student-4th Grade” that was categorized as “not in labor force-student” and “student related-other”.

Examples from the “freq. ≤ 5” dataset include “snowplow driver, road maintenance” and “legal assistant, retired 2012”. Both entries were placed under several categories. The first entry was categorized as “name of
occupation”, “type of occupation” and “industry” with quality issues of “misspelling” and “other-punctuation mark”. The second entry was categorized as “name of occupation”, “type of occupation”, “not in labor force-other” and “miscellaneous-date” with quality issue of “other-punctuation mark”.

After completing the manual categorization of unique occupation entries in both datasets (n=2,717), counts and percentages for both the “unique” and “overall” entries were calculated for the frequencies of “unique” occupation entries occurring within the dataset. “Overall” entries within a specific frequency group represent the overall number of times the corresponding “unique” entry occurred in the dataset.

Table 3: Categorization Schema for “Data Quality” Issues

<table>
<thead>
<tr>
<th>#</th>
<th>Issue</th>
<th>Brief Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Misspelling</td>
<td>Entry includes misspelling</td>
<td>Message therapist, software engineer</td>
</tr>
<tr>
<td>2</td>
<td>Acronym/Abbreviation</td>
<td>Entry includes acronym or abbreviation</td>
<td>CAN, Administrative asst.</td>
</tr>
<tr>
<td>3</td>
<td>Ambiguous</td>
<td>Entry includes ambiguous information</td>
<td>Account, Domestic Goddess</td>
</tr>
<tr>
<td>4</td>
<td>Multiple Entries</td>
<td>Entry includes two or more distinct terms that represent more than one category or two occupations divided by “/”</td>
<td>Student/waitress, owner/operator</td>
</tr>
<tr>
<td>5</td>
<td>Other</td>
<td>Entry is informal use of word, all in uppercase letters, includes punctuation marks such as ‘.’ or ‘,’ etc. Provide details as part of comment/notes (e.g., “uppercase” or “punctuation mark”).</td>
<td>MANAGER, Society, quit</td>
</tr>
</tbody>
</table>

Results

Table 4 depicts the frequency distribution of unique and overall entries, and illustrates the selection of a frequency of “5” as the cutoff for the two datasets for performing data analysis. The table provides a summary of the “unique” occupation entries (n=53,424), which includes the overall and unique counts and percentages of occupation entries grouped into the frequency of the occurrence of a specific “unique” entry within the dataset. “Frequency” represents the number of times a specific “unique” entry occurred in the dataset. “Unique” occupation entries that occurred more than 10 times within the dataset have been grouped into the “>10” frequency”.

Table 4: Frequency Distributions for Unique and Overall Occupation Entries

<table>
<thead>
<tr>
<th>Frequency</th>
<th># Unique Entries</th>
<th>% Unique</th>
<th># Overall Entries</th>
<th>% Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;10</td>
<td>1,319</td>
<td>2.5%</td>
<td>116,190</td>
<td>62.5%</td>
</tr>
<tr>
<td>10</td>
<td>121</td>
<td>0.2%</td>
<td>1,210</td>
<td>0.7%</td>
</tr>
<tr>
<td>9</td>
<td>141</td>
<td>0.3%</td>
<td>1,269</td>
<td>0.7%</td>
</tr>
<tr>
<td>8</td>
<td>176</td>
<td>0.3%</td>
<td>1,408</td>
<td>0.8%</td>
</tr>
<tr>
<td>7</td>
<td>238</td>
<td>0.5%</td>
<td>1,666</td>
<td>0.9%</td>
</tr>
<tr>
<td>6</td>
<td>341</td>
<td>0.6%</td>
<td>2,046</td>
<td>1.1%</td>
</tr>
<tr>
<td>5</td>
<td>484</td>
<td>0.9%</td>
<td>2,420</td>
<td>1.3%</td>
</tr>
<tr>
<td>4</td>
<td>756</td>
<td>1.4%</td>
<td>3,024</td>
<td>1.6%</td>
</tr>
<tr>
<td>3</td>
<td>1,397</td>
<td>2.6%</td>
<td>4,191</td>
<td>2.3%</td>
</tr>
<tr>
<td>2</td>
<td>4,016</td>
<td>7.5%</td>
<td>8,032</td>
<td>4.3%</td>
</tr>
<tr>
<td>1</td>
<td>44,435</td>
<td>83.2%</td>
<td>44,435</td>
<td>24.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>53,424</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>185,891</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>

Figure 2 compares the “overall” percentages of the “contents” of the unique occupation field distributed among the five main content groups between the two datasets. Noticeably, around 22% of the overall data being analyzed (in both datasets) were indicated as “not in labor force”. The top five “overall” entries categorized as “not in labor force”, regardless of identified data quality issues, were student (n=10,555), homemaker (n=4,693), “housewife” (n=958), retired (n=679) and stay at home (n=541). Results of the “overall” counts of the three
subcategories associated with the category “not in labor force - student” were: “student related – status” (n=31), “student related – type” (n=590) and “student related – other” (n=716).

Figure 2: Distribution of the “Overall” Percent of Entries across Main Categorization Groups for “Contents”

Applying the categorization schemas to identify the contents and data quality issues of the two datasets being analyzed are summarized in Tables 5 and 6. The results are presented to show the count and percentages of both the “unique” and “overall” occupation entries between the two datasets. Percent of “unique” entries categorized as “name of occupation” (64.9%) was higher in the “freq. > 5” dataset. As anticipated, the percent of “unique” entries categorized as “miscellaneous” was much higher in the “freq. ≤ 5” dataset (23.9%). Examples of “miscellaneous” entries included “odd jobs”, “infant”, “data” and “grocery”. The “freq. ≤ 5” dataset was also identified as having a higher percentage of quality issues indicated as “other” such as punctuation marks in the entry or the entire entry in uppercase letters.

Table 5. Categorization of Contents for Occupation Entries

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>“freq. &gt; 5” Dataset (n=2,336)</th>
<th></th>
<th></th>
<th>“freq. ≤ 5” Dataset (n=381)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Count (%)</td>
<td></td>
<td>Count (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Name of Occupation</td>
<td>1,515 (64.9%)</td>
<td></td>
<td>215 (56.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Type of Occupation</td>
<td>1,072 (45.9%)</td>
<td>212 (53.7%)</td>
<td>212 (57.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Industry</td>
<td>234 (10.0%)</td>
<td>27 (7.1%)</td>
<td>27 (7.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Workplace</td>
<td>116 (5.0%)</td>
<td>66 (17.3%)</td>
<td>66 (16.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Job Duties</td>
<td>164 (7.0%)</td>
<td>43 (11.3%)</td>
<td>43 (9.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Employer Name</td>
<td>88 (3.8%)</td>
<td>57 (15.0%)</td>
<td>57 (14.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Equipment</td>
<td>22 (0.9%)</td>
<td>2 (0.5%)</td>
<td>2 (0.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Employment Status</td>
<td>13 (0.6%)</td>
<td>12 (3.1%)</td>
<td>12 (2.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Unemployment Status</td>
<td>16 (0.7%)</td>
<td>3 (0.8%)</td>
<td>3 (0.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Not in Labor Force</td>
<td>36 (1.5%)</td>
<td>9 (2.4%)</td>
<td>9 (2.4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Student Related -Status</td>
<td>2 (0.1%)</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.2</td>
<td>Student Related -Type</td>
<td>22 (0.9%)</td>
<td>5 (1.3%)</td>
<td>5 (1.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.3</td>
<td>Student Related -Other</td>
<td>41 (1.8%)</td>
<td>7 (1.8%)</td>
<td>7 (1.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Not in Labor Force -Others</td>
<td>73 (3.1%)</td>
<td>21 (5.5%)</td>
<td>21 (4.6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>NA/None</td>
<td>3 (0.1%)</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Miscellaneous</td>
<td>76 (3.2%)</td>
<td>91 (23.9%)</td>
<td>91 (22.1%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Significant quality issues were found in both of the study datasets. When looking at the number of quality issues identified within a single “unique” entry, it was found that in the “freq. ≤ 5” dataset, there were about 21% “unique” occupation entries that had one quality issue identified, 18% with two quality issues, 11% with three quality issues and 1% with four identified quality issues. Comparing these findings with the quality issues identified in the “freq. > 5” dataset, it was found that 9% had one quality issue, 12% with two quality issues, 1% with three quality issues and 1% with four identified quality issues.

**Discussion**

The overall goal of this study was to understand current documentation practices for occupation in the EHR for informing efforts to structure and standardize this information for subsequent use. Although, when developing the categorization schema, federal laws related to occupation were not considered (e.g., Fair Labor Standards Act), our analysis demonstrates that data being documented in the free-text occupation field under the socioeconomic section of the EHR being studied partially reflect current standards and recommended formats from reports that have been published. As per the proposed NIOSH representation as well as other national reports and recommendations, occupation and occupation details can be captured in the following categories: (1) occupation, (2) industry, (3) employment status, (4) employer, (5) work schedule, (6) occupational injury, (7) occupational exposures and (8) work relatedness. Preliminary findings from a related study focused on evaluating the adequacy of the ODH Model developed by NIOSH supported the robustness of the ODH Model for representation of occupational information. Designing the EHR to capture occupational information for a patient in a free-text field resulted primarily in concepts related to occupation, status, industry, and employer. Having the data stored in a free-text field also creates a chance for errors during the data entry process, which could, in turn, result in lower quality of occupation data for secondary applications like research and population health interventions.

Four main observations resulted from our analysis. Within the study dataset, 38% (n=114,236) of the entries had the number “0” entered in the occupation field or had no entry at all. These were considered missing values and thus were not included in the primary analysis. Another observation was that 14% (n=44,435) had an entry that only occurred once in the entire dataset. This was mostly due to user-created terms, errors or descriptive statements being entered in the field. Examples of such descriptive sentences are “homemaker, former rn, 13 kids, 2 biologic” and “fishing guide in Ontario during summers, retired science teacher”. Reasons for this issue could be due to the fact that the occupation field is a free-text field, lack of awareness or training on the part of users on entering this information, or the absence of appropriate fields designed for structured documentation of associated occupational information. The third observation was related to the frequencies of the unique occupation entries. As the frequencies of the entries decreased, the relative number of identified quality issues increased. Therefore, the decision to split the data into two separate datasets based on the frequencies of the unique occupations was made for better overall management and analysis. The fourth observation was related to entries that described individuals not in the workforce or student-related entries such as “retired”, “on disability”, “stay at home mom”, “housewife” and “4th grade student”. Subject matter experts were faced with finding the most appropriate category to represent this group within the coding schema, as there were not any available standards that clearly addressed this group in a consistent manner. A recommendation from this is that there should be a separate field or set of fields in the EHR that represent these types of individuals and associated information. Also essential is the need to develop standards and measures that would address this specific issue and design the EHR to capture information for these individuals whom are not in the labor force.

The results of this study also highlighted significant data quality issues associated with this data field and the relative low utility of this data for secondary purposes such as research, policy and population initiatives. Due to the fact that the study only analyzed one free-text field (Occupation) and represented data from a single EHR in one healthcare system (Epic EHR at Fairview Health Services), these findings may not be generalizable or represent the quality of occupation documentation in other EHR systems or other clinical settings. Since Fairview Health Services includes six different hospitals (four are metropolitan based and two are rural), an
academic practice with quaternary care, and a community practice with a specialized children’s hospital, the
initial findings potentially represent a breadth of documentation practices for occupational information. In
addition, the methods for characterizing the contents and quality of this information could be adapted and
applied at other healthcare systems. Future studies will incorporate federal laws related to occupation, normalize
the “occupation” field entries, group and aggregate synonym entries, and analyze the occupation field in relation
to other associated fields in the social history module (i.e., Employer and Comments as shown in Table 1), as
well as other parts of the EHR that may include occupational information (e.g., clinical notes). More formal data
quality assessments, such as those described by Weiskopf and Weng17, will also be needed.

Examining a single EHR that reflects a standard input mechanism by a particular vendor may result in innate
limitations in how occupational data are currently being captured. Broader implications of this work include
informing improved EHR interfaces for capturing occupational data and how to codify free-text occupation
information stored in the EHR. This study also helps set additional foundation for efforts in the area of NLP to
analyze free-text stored in the EHR related to the occupational history of patients. NLP techniques can be used
to extract, structure, and encode relevant information from free-text data for subsequent use. Knowing what the
target is, in this case occupational information, and the type of model used to extract from text will set the
foundation for mapping to standardized terminologies such as SOC13, NAICS15, and SNOMED-CT14 that
include codes for occupation and industry. For example, the social context hierarchy in SNOMED-CT14 is
designed to cover social conditions and circumstances significant to healthcare that includes a sub-hierarchy for
occupation, which could be used to standardize different values found from this study. Future work could
involve developing and evaluating NLP techniques for the different categories of content in the occupation field
as identified in this study.

Conclusion

With the increased adoption of EHR systems and the growing recognition of social factors in impacting health
outcomes, there is a need to understand the current status of the information being stored and captured in EHRs
to increase the value of information that can be obtained. This study involved performing a content analysis of
data from a free-text occupation field of an EHR over a selected time period by categorizing the contents of
information being captured and identifying associated data quality issues, using developed categorization
schemas. The findings of this study have implications in terms of system design, user training, and
implementation of relevant standards, including vocabulary related to occupation and industry.

Acknowledgements

This study was supported by the grant from the National Library of Medicine #R01LM011364 and the
University of Minnesota Clinical Translational Science Institute #8UL1TR000114.

References

1. Office of the National Coordinator for Health IT (ONC). Health IT Dashboard [Internet]. [updated 2015;
http://www.nap.edu/catalog.php?record_id=13207
3. National Institute for Occupational Safety and Health (NIOSH). Structuring Patient Work Information in
EHRs to Improve Patient Care and Public Health; Occupational Data for Health (ODH) Model. Public
Health Informatics Conference; Atlanta, GA. 2014.
4. Health Level Seven (HL7) - IHE PCC Technical Committee. Integrating the Healthcare Enterprise (IHE)
Patient Care Coordination (PCC) Technical Framework Supplement - CDA Content Modules; Trial
Implementation Guide [Internet]. 2014 Dec [cited 2016 June 22]. Available from:
5. Institute of Medicine (IOM). Capturing Social and Behavioral Domains in Electronic Health Records:
http://www.nationalacademies.org/hmd/Reports/2014/Capturing-Social-and-Behavioral-Domains-in-
Electronic-Health-Records-Phase-1.aspx
6. Institute of Medicine (IOM). Capturing Social and Behavioral Domains and Measures in Electronic Health
Records, Phase 2. Washington, DC [Internet]. 2014 [cited 2016 June 22]. Available from:
health-records

1715
SMASH: A Data-driven Informatics Method to Assist Experts in Characterizing Semantic Heterogeneity among Data Elements

William Brown III, DrPH, MPhil, MA1,2, Chunhua Weng, PhD1, David K. Vawdrey, PhD1,3, Alex Carballo-Diéguez, PhD2, Suzanne Bakken, PhD, RN1,4

1Department of Biomedical Informatics, Columbia University, New York, NY; 2HIV Center for Clinical and Behavioral Studies, NY State Psychiatric Institute & Columbia University, New York, NY; 3New York-Presbyterian Hospital Value Institute, New York, NY; 4School of Nursing, Columbia University, New York, NY

Abstract

Semantic heterogeneity (SH) is detrimental to data interoperability and integration in healthcare. Assessing SH is difficult, yet fundamental to addressing the problem. Using expert-based and data-driven methods we assessed SH among HIV-associated data elements (DEs). Using Clinicaltrials.gov, we identified and obtained eight data dictionaries, and created a DE inventory. We vectorized DEs by study, and developed a new method, String Metric-assisted Assessment of Semantic Heterogeneity (SMASH), to find DEs: similar in A and Bn, unique to An, and unique to Bn. An HIV expert assessed pairs for semantic equivalence. Heterogeneous DEs were either semantically-equivalent/syntactically-different (HIV-positive/HIV+/Seropositive), or syntactically-equivalent/semantically-different (“Partner”[sexual]/“Partner”[relationship]). Context of usage was considered. SMASH aided identification of SH. Of 1,175 DE from pairs, 1,048 (87%) were semantically heterogeneous and 127 (13%) were homogeneous. Most heterogeneous pairs (97%) were semantically-equivalent/syntactically-different. Expert-based and data-driven methods are complementary for assessing SH, especially among semantically-equivalent/syntactically-different DE. Similar expert-based/data-driven solutions are recommended for resolving SH.

Introduction

Interoperability of health information systems is critical to the effective provision of health care. Interoperability involves the exchange and interchange of diagnostic, treatment, prescription, and billing-related data. For these information types to be exchanged accurately and effectively, healthcare providers working with these data must understand the true meaning, or semantics, of the data and their relationship to each other1–3. Moreover, integration of diverse domain-associated datasets provides new opportunity for analysis, the development of new knowledge, improved disease discovery, diagnoses, treatment, and intervention4. It also provides an opportunity to resolve common data analytic problems, such as small sample size, need for subgroup analysis, need for comparative analyses between populations, and low statistical power2,4. For these reasons, integration of health-related data becomes critical to health care. Semantics play an instrumental role in the integration process, assuring that researchers who use the integrated datasets are analyzing what they truly mean to analyze4,5.

The difference in meaning and interpretation of data elements (DEs) is known as semantic heterogeneity (SH)6,7. It can occur when database schemas or datasets for the same domain are developed independently6,8. Because of the critical relationship of semantics to health-related data, SH is a major source of challenges for data integration, lack of interoperability among health information systems, and barriers to the provision of accurate and effective health care, and knowledge generation9–11. For this reason, resolving SH, also known as semantic harmonization, is key to achieving both interoperability among healthcare systems and integration of diverse domain-associated datasets11.

SH results from independent development of database schema or datasets for the same domain, forming differences in meaning and interpretation of DEs and data values6,9. Decomposing the various sources of SH provides a basis for understanding how to map and transform data to overcome these differences. Unfortunately, the nature or extent of SH in a given domain is rarely known or quantified. Most researchers try to get around the issue of SH quickly by immediately identifying semantically homogeneous DEs12,13. Consequently, the optimal solution may be missed, the issue of SH remains, and to what extent and why is left unknown. If SH is to be addressed in a meaningful and long-lasting way, we must characterize the SH. Thus, assessing how much SH exists among domain-associated DEs is an important, yet challenging informatics task.

Data-driven methods such as approximate string matching are relatively high-throughput methods, but only work at the syntactic level and are weak at dealing with semantics. Expert-based methods support consensus semantic
reasoning between expert raters and across multidisciplinary domain-similar groups; however, these methods are often time-consuming and expensive\textsuperscript{6,13,14}. In this research, we used both expert-based and data-driven informatics methods to assess SH among empirically generated DEs in HIV-associated data dictionaries. Specifically, we present a new method, String Metric-assisted Assessment of Semantic Heterogeneity (SMASH), to find similar DEs and assist experts in assessing DE pairs for semantic equivalence.

**Background**

SH is one of the more important and challenging sources of differences in heterogeneous datasets\textsuperscript{11}. Moreover, SH is compounded by the flexibility of semi-structured data and various coding methods applied to unstructured data, such as narrative documents\textsuperscript{13,15}. Such is the case within the HIV research domain.

Since the early years of the epidemic, National Institutes of Health (NIH) Institutes and Centers, as well as other funding agencies, have supported numerous studies that have collected longitudinal data on HIV-associated variables and HIV-disease specific measures (e.g., viral load and CD4 counts)\textsuperscript{16}. Most of these studies research comparable populations and include patients with similar demographic characteristics. However, assessment approaches, instruments, and protocols are often created independently\textsuperscript{13,16}. Given the complexity and multitude of possibly interrelated factors that influence HIV outcomes, larger, integrated datasets could contribute significantly to advanced analysis.

Because HIV research is often conducted independently by a variety of clinical trials networks and investigators, schematic and syntactic differences are generated that contribute to SH. What’s more, few efforts have been made to combine data across studies\textsuperscript{13,16,17}. However, such harmonization would significantly increase the value of the collected data to address HIV-related questions for those living with, or at risk of acquiring, HIV. For these reasons, we focused our efforts on characterizing the SH of the datasets in the HIV research domain.

**Methods**

**Empirically Generated Data Sources**

We set out to characterize SH among empirically generated HIV-associated research DE. In this study, empirically generated data sources were defined as those DE that derived from the practice of HIV clinical or behavioral research. To this end, HIV-associated research data dictionaries were collected. This study was reviewed by the institutional review board and declared to be exempt.

**Data-driven Methods**

We systematically searched Clinicaltrials.gov for relevant HIV-associated studies. We used fifteen NIH criteria outlined in RFA-MH-14-200 “Integration and Analysis of Diverse HIV-Associated Data (R03)”. Criteria for DEs included: HIV+, HIV-, Male, Female, Transgender, men who have sex with men (MSM), Adult, Youth, Children, Neuro-psychological, Psychosocial, Behavioral, Biological/Biomarkers, Longitudinal, accessible data. After identifying eligible research studies with datasets that included several criteria, we approached principal investigators (PIs) of the research studies via email using a participation recruitment letter. The letter summarized the nature of the study, outlined the use of the data, and described the approvals needed for the use of the data dictionaries.

Once we obtained data dictionaries, we extracted DEs (both variables and metadata) from the data dictionaries. We used Excel to create an inventory of DEs, organized by study. Data elements for each study were vectorized using R which requires that vectors must be of the same data type (e.g. character, logical, numeric, integer). Thus, all non-character string DEs were converted (“coerced”) to character strings so that they could be compared (see Figure 1).
Next, to begin the DE comparative analysis, we used R approximate (Fuzzy) string matching to compare the DE vectors\textsuperscript{18,19}. We did this by computing distance metrics (in this case, the syntactic difference) between DE pairs. We compared DE vectors in the following ways (Figure 2) and asked the following questions of the data:

- What are the similar DEs in study $A_{n,s}$ and study $B_{n,s}$?
- What are the unique DEs in study $A_{n,s}$, but not in study $B_{n,s}$?
- What are the unique DEs in study $B_{n,s}$, but not in study $A_{n,s}$?

To find similar DEs and assist experts in assessing DE pairs for semantic equivalence we devised SMASH, a string metric-assisted assessment of semantic heterogeneity. SMASH assumes a possible relationship between syntactic and semantic similarity, and uses one or more string metrics to compare syntactic similarity of DEs to find DEs that may also be semantically similar. String comparison, similarity and distance, metrics used for SMASH were provided by the \textit{stringdist} package (R library(stringdist)) in the R statistical programming environment\textsuperscript{18,19}. We used the \textit{stringsim} command to compute pairwise string similarities between elements of character string vectors in vector $A_n$ and $B_n$, where the vector with fewer elements is recycled. This involved first calculating the distance using \textit{stringdist}, set to Levenshtein distance (LD) (method='lv'). The \textit{stringsim} command then divides the distance by the maximum achievable distance and subtracts the value from 1. For \textit{stringsim} we set the method to Jaro-Winker distance (JWD) (method='jw'). This produces a score between 0 – 1 (1 = complete similarity [distance 0] or 0 = complete dissimilarity)\textsuperscript{18}. The function then returns a vector with similarity scores (see Figure 3). Recursion was used to iterate through all of the possible pairs between vectors. These methods are case sensitive; thus, all character strings were changed to lowercase before the analysis.
Expert-based Methods

To continue the DE comparative analysis, an HIV expert assessed all vector pairs according to DE distance and similarity scores. DEs were first organized in their comparison category: (A, B), (A, not B), and (B, not A). Pairs that were similar in A∞ and B∞ were assessed first, followed by data elements that were unique to A∞, and unique to B∞. Among (A, B) pairs, the expert first assessed those pairs with a distance of less than 13, since these commonly included simpler strings that could be assessed quickly; however, all (A, B) pairs were eventually assessed. All groups of pairs were organized from highest similarity score to lowest similarity score. Context of usage was also factored into judgment of SH. The expert also confirmed that unique DEs were truly unique (i.e., having no semantic or syntactic equivalence to a DE in the vector to which it was compared).

To identify semantically heterogeneous DEs, two types of DE pairs were assessed, either semantically-equivalent/syntactically-different DEs (e.g., HIV-positive/HIV+/Seropositive), or syntactically-equivalent/semantically-different DEs (e.g., “Partner”[sexual] vs “Partner”[relationship]). Although unique DEs contribute to heterogeneity in general, they may not directly contribute to the SH heterogeneity between two domain-associated datasets. To find the percent of data elements that specifically contributed to SH, we subtracted the number of unique DE—((A, not B) and (B, not A))—from the total number of DE. This left only those DEs with semantic similarity, syntactic similarity, or both. To complete our analysis, we divided the number of semantically heterogeneous DEs over all non-unique DEs. This can be represented by (A||B, SH)/(A, B) – ((A, not B) + (B, not A).

Results

Clinicaltrials.gov Search

Our search from Clinicaltrials.gov returned 453 studies. Principal Investigators (PIs) with an existing relationship with either this study’s PI, institution, or internal review board (IRB) were contacted. This was done to help facilitate not only the procurement of data dictionaries, but also the acquisition of related datasets for use in subsequent aims of this research, which will include use of actual participant data. We sent invitations to the PIs of 20 studies and received responses from ten PIs, and of the ten PIs we were able to confirm that 18 datasets were available for use in our study. The total number of participants represented by the 18 datasets was >4,300. Currently, of the 18 that were indicated as possible to use in our study, we have procured and analyzed eight (see Tables 1-3), six are still being procured, and four could no longer be offered for analysis. The total number of DE among the eight studies we have analyzed was 1,142. Examples of DE from those eight studies are provided in Table 4 and the number of DE per study is indicated in Table 5.
Table 1. Descriptions of included research studies.

<table>
<thead>
<tr>
<th>Name of Study</th>
<th>Funder</th>
<th>Study Period</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Gel: Microbicide Safety and Acceptability in Young Men</td>
<td>NICHD/NIMH: 5R01HD059533 / NIH: R01 HD059533</td>
<td>2010 - 2013</td>
<td>GEL</td>
</tr>
<tr>
<td>Combination HIV Antiretroviral Rectal Microbicide (CHARM) project</td>
<td>NIAID: 5U19AI082637</td>
<td>2009 - 2014</td>
<td>CHARM</td>
</tr>
<tr>
<td>The Options Now study</td>
<td>Bill and Melinda Gate Foundation</td>
<td>2012 - Ongoing</td>
<td>ON</td>
</tr>
<tr>
<td>Rapid HIV Home Test and Decision-making among HIV-negative Men</td>
<td>NIH: R01 MH079692</td>
<td>2009 - 2013</td>
<td>HT2</td>
</tr>
<tr>
<td>Topical Microbicide Acceptability</td>
<td>NIH: R01 HD046060</td>
<td>2003 - 2007</td>
<td>Top Mic</td>
</tr>
<tr>
<td>Structural Intervention to Integrate Sexual and Reproductive Health Information into HIV Care</td>
<td>NIMH, R01; 2006-2011</td>
<td>2006 - 2011</td>
<td>SI</td>
</tr>
<tr>
<td>Pathways to Engagement in HIV Care for Newly-Diagnosed South Africans</td>
<td>NIMH: NIMH R01-MH08356</td>
<td>2009 - 2014</td>
<td>Pathways</td>
</tr>
<tr>
<td>Use of Design Science for Informing the Development of a Mobile App for Persons Living with HIV</td>
<td>CDC: 1U01PS00371501</td>
<td>June 2013 - Oct 2013</td>
<td>Design Science</td>
</tr>
</tbody>
</table>

The eight data dictionaries procured met from 63% to 100% of the population inclusion criteria outlined in RFA-MH-14-200 “Integration and Analysis of Diverse HIV-Associated Data (R03)”, and all included data involving youth (18-25) and adults (ages 26-55). None included children (ages 12-17), which was one of the originally desired criteria in RFA-MH-14-200, but ultimately dropped in our study for IRB associated reasons. Six included HIV-positive participants, six included HIV-negative participants, and four included a combination of both. All eight also included male participants, most of which were MSM except for one study. Six studies included female participants, and three studies included transgender participants (see Table 2).

Table 2. Data populations by study.

<table>
<thead>
<tr>
<th>Study</th>
<th>HIV+ Male</th>
<th>HIV- Male</th>
<th>Female</th>
<th>Adult (26-55)</th>
<th>Youth (18-25)</th>
<th>MSM</th>
<th>Trans</th>
<th>IC%*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEL</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>88%</td>
</tr>
<tr>
<td>CHARM</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
</tr>
<tr>
<td>ON</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>88%</td>
</tr>
<tr>
<td>HT2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>63%</td>
</tr>
<tr>
<td>Top Mic</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>75%</td>
</tr>
<tr>
<td>SI</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>63%</td>
</tr>
<tr>
<td>Pathways</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>75%</td>
</tr>
<tr>
<td>Design Science</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>100%</td>
</tr>
</tbody>
</table>

Study abbreviations are listed in Table 1. *Percentage of population inclusion criteria met by the data dictionary.

Similarly, the data dictionaries procured met from 60% to 100% of the data type inclusion criteria. The number of studies that contain specific data types was as follows: Behavioral (8), Longitudinal data (8), Psychosocial (7), Neuro-psychological (5), and Biological/Biomarkers (5) (see Table 3).
Table 3. Data types by study.

<table>
<thead>
<tr>
<th>Study</th>
<th>Neuro-psychological</th>
<th>Psychosocial</th>
<th>Behavioral</th>
<th>Biological/Biomarkers</th>
<th>Longitudinal</th>
<th>IC%*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEL</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>80%</td>
</tr>
<tr>
<td>CHARM</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>80%</td>
</tr>
<tr>
<td>ON</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
</tr>
<tr>
<td>HT2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
</tr>
<tr>
<td>Top Mic</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
</tr>
<tr>
<td>SI</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>Pathways</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
<td>60%</td>
</tr>
<tr>
<td>Design Science</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>100%</td>
<td>60%</td>
</tr>
</tbody>
</table>

Study abbreviations are listed in Table 1. *Percentage of data type inclusion criteria met by the data dictionary.

SMASH and Comparative Similarity Analysis

Tables 4 and 5 below show the results from the data-driven, expert-based similarity analysis. Table 4 shows the results from SMASH, the data-driven string metrics described in the “Data-driven Methods” section, and Table 5 shows the results from the expert-based assessment of the pairs found by using SMASH (described in the “Expert-based Methods” section). There were a total of 1,175 pairs of semantic similarity DEs. We found a total of 1,048 (87%) cases of SH similarity between study vectors A_n and study B_n. Most heterogeneous pairs (97%) were semantically-equivalent/syntactically-different. Of the 1,048 >50% were repeated matches with two or more studies. Conversely, 127 (13%) pairs were semantically homogeneous. Semantically homogeneous pairs most often consisted of demographic DEs (e.g. race, age, gender, sex) and administrative DEs (data, time, consent, testing). Studies with a smaller number of DEs that were being compared to studies with a larger number of DEs had the most similarity overlap of their DEs. That is, smaller studies’ DEs most often matched to at least one of the DEs of a larger study.

Table 4. Sample of data elements, comparative analysis, and identification of semantically heterogeneous DE using SMASH.

<table>
<thead>
<tr>
<th></th>
<th>Levenshtein Distance</th>
<th>Jaro Winler</th>
<th></th>
<th>Levenshtein Distance</th>
<th>Jaro Winler</th>
</tr>
</thead>
<tbody>
<tr>
<td>(hiv+, hiv test)</td>
<td>5</td>
<td>0.8</td>
<td>(hiv status, hiv test)</td>
<td>5</td>
<td>0.89</td>
</tr>
<tr>
<td>(hiv+, hiv status)</td>
<td>7</td>
<td>0.78</td>
<td>(hiv status, hiv positive)</td>
<td>7</td>
<td>0.82</td>
</tr>
<tr>
<td>(hiv+, seropositive)</td>
<td>10</td>
<td>0</td>
<td>(hiv status, hiv knowledge)</td>
<td>9</td>
<td>0.74</td>
</tr>
<tr>
<td>(hiv+, hiv knowledge)</td>
<td>10</td>
<td>0.76</td>
<td>(hiv status, crystal meth)</td>
<td>10</td>
<td>0.51</td>
</tr>
<tr>
<td>(hiv+, methamphetamine)</td>
<td>13</td>
<td>0</td>
<td>(hiv status, sexual partner)</td>
<td>12</td>
<td>0.49</td>
</tr>
<tr>
<td>(hiv+, status of partner)</td>
<td>17</td>
<td>0</td>
<td>(hiv status, sero-discordant)</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>(hiv+, black (non-hispanic))</td>
<td>18</td>
<td>0</td>
<td>(hiv status, african american/ black)</td>
<td>19</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Study abbreviations are listed in Table 1. Number of DE per study are in parenthesis.

We identified a total of 498 SH DEs without overlap between studies. Heterogeneous DEs were either semantically-equivalent/syntactically-different (ex. HIV positive = HIV+ = Seropositive), or syntactically-equivalent/semantically-different (e.g., “Partner” [sexual] vs. “Partner” [relationship]). Context of usage was considered. Semantically-equivalent/syntactically-different were the most common pairs identified (Additional examples can be found in Table 5). Finally, we calculated the percent of heterogeneous pairs over all pairs from each vector comparison. Word distance varied (min=0, max=23, avg=7). We were able to calculate that the average amount of SH between vector pairs was 87% from the results in Table 6. The fact that most heterogeneous pairs (97%) were semantically-equivalent/syntactically-different did not change in this part of the analysis.

Table 5. Number of similar DE pairs between An and Bn.

<table>
<thead>
<tr>
<th>Bn</th>
<th>An</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CHARM</td>
</tr>
<tr>
<td></td>
<td>(439)</td>
</tr>
<tr>
<td>CHARM (439)</td>
<td>42</td>
</tr>
<tr>
<td>ON (78)</td>
<td>42</td>
</tr>
<tr>
<td>Pathways (24)</td>
<td>24</td>
</tr>
<tr>
<td>GEL (68)</td>
<td>28</td>
</tr>
<tr>
<td>HT2 (45)</td>
<td>32</td>
</tr>
<tr>
<td>Top Mic (76)</td>
<td>45</td>
</tr>
<tr>
<td>SI (237)</td>
<td>141</td>
</tr>
<tr>
<td>Design Science (175)</td>
<td>108</td>
</tr>
</tbody>
</table>

Table 6. Semantic heterogeneity among DE vector pairs by study.

<table>
<thead>
<tr>
<th>Bn</th>
<th>An</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CHARM</td>
</tr>
<tr>
<td></td>
<td>(439)</td>
</tr>
<tr>
<td>CHARM (439)</td>
<td>95%</td>
</tr>
<tr>
<td>ON (78)</td>
<td>92%</td>
</tr>
<tr>
<td>Pathways (24)</td>
<td>89%</td>
</tr>
<tr>
<td>GEL (68)</td>
<td>97%</td>
</tr>
<tr>
<td>HT2 (45)</td>
<td>98%</td>
</tr>
<tr>
<td>Top Mic (76)</td>
<td>87%</td>
</tr>
<tr>
<td>SI (237)</td>
<td>87%</td>
</tr>
<tr>
<td>Design Science (175)</td>
<td>91%</td>
</tr>
</tbody>
</table>

Study abbreviations are listed in Table 1. Number of DE per study are in parenthesis.

Discussion

In this study we used a combination of data-driven (e.g., SMASH) and expert-based informatics methods to characterize SH. By using SMASH to analyze DE from data dictionaries from empirically generated sources, we gained a better understanding of the nature and extent of SH in the HIV research domain. Our use of empirically developed data is unique in that we are not responding to an internal need to address SH in our own system, but rather to understand the problem within a domain of research. Most often, attempts to address SH have been in reaction to a need for an institution to integrate its data with various internal systems or with the data of other institutions. As a
result, they attempt to harmonize the data without characterizing the extent of the problem. We started with first assessing the problem and used diverse empirically generated sources of data. This gave us the advantage of understanding different relationships between the nature (e.g., semantically-equivalent/syntactically-different) of the data and the extent of SH.

For instance, we found that although all of the data dictionaries collected had more than 60% of overlapping populations and data types, they still had a very high level of SH among their DEs. However, we also found that the crossover between research teams also had an impact on the amount of SH between different studies. For instance, though HT2, GEL, SI, and Top Mic had different PIs, they all had overlapping research staff. As a result, their percent of SH was notably lower than the other studies. Also, though GEL and Top Mic were conducted at different institutions, they had at least one PI (or Co-PI) in common. In contrast, Pathways was conducted at the same institution as GEL and HT2, and had no staff in common. In this case, there was still a high amount of SH between Pathways and the other two studies (GEL and HT2).

The use of Clinicaltrials.gov was instrumental to the dataset identification process. Furthermore, using both our data-driven SMASH method and expert-based methods in our similarity analysis had distinct benefits over using one method alone. The SMASH method for the discovery of syntactically similar DEs helped to identify DEs that potentially were semantically similar. This sped up the process of discovery. Also critical was leveraging both the distance score and the similarity score in the organizational process. We found that by leveraging both, those DEs that had a low distance score and a high similarity score often had a notably close relationship. This method is not only helpful for identification of SH but also may facilitate the identification of common DEs. However, we also found that data-driven methods alone were insufficient.

The expert-based methods we used (see Figures 1-3) were instrumental in identifying those DE that were semantically similar, but completely syntactically different (i.e. had a high distance score and a similarity score of 0). For example, HIV+ and seropositive, have a direct 1:1 semantic relationship; however, they have a distance of 10 and a similarity score of 0 based on SMASH (see Table 4). The HIV expert was able to find this relationship, but the data-driven methods alone may have missed it. Though semantic learning may have done a better job, computational methods require additional data to learn how syntactically different relationships are semantically similar. This level of data is not often present in most data dictionaries. Similarly, HIV status and sero-discordant have a close semantic relationship, but these also had a high distance (14) and a similarity score of 0. Lastly, one of the fundamental characteristics about the heterogeneity that we found was the type of SH. We found that most heterogeneous pairs were semantically-equivalent/syntactically-different. This has strong implications for how, or on what, to focus attempts to resolve SH.

**Future Work**

Future development of methods to resolve SH, particularly those in the HIV research domain, should focus on variables that are semantically-equivalent/syntactically-different, which we found most likely to reflect SH. Variable types that we found to be commonly homogeneous, such as demographics, should have lower priority. Also, Clinicaltrials.gov should be leveraged for identifying empirically generated data sources for future SH discovery in other research domains. Future work should also include data dictionaries on children and transgender individuals. It should also focus on data dictionaries from international research. Moreover, looking at data dictionaries alone may limit the data-driven approach. Future work should also look at actual values for each DE to see if any additional information about the DE can be gained from the data it elicits. This could make for a stronger data driven approach. Methods to identify SH and achieve semantic harmonization should be developed in a way that more seamlessly integrates both data-driven and expert-based informatics methods for identifying SH. SMASH should also be systematically evaluated against other computational methods and assessed for sensitivity and specificity. Lastly, future work should also aim to analyze a larger number DEs from basic science research (e.g., genomics).

**Limitations**

The relatively small number of empirically derived DE sources is a limitation. Moreover, the search parameters used in Clinicaltrials.gov did not account for the full breadth of topics in the HIV research domain (e.g., genomics, proteomics, virological research). Also, DEs that address specific variables related to some sub-populations were lacking in the data dictionaries (i.e., children, transgender individuals).
Conclusion

SH is detrimental to interoperability between healthcare systems and data integration among domain-associated DEs in health research. For multiple data sources to interoperate with one another, it is essential to reconcile these semantic differences. Moreover, SH has the potential to negatively impact the interpretation of NIH funded study results, and health policy implications that are drawn from the analyses of data from various studies. As a result, SH may also have an effect on FDA medical product approvals and the comparability of results across trials. Thus, it is imperative that health policy makers encourage methods to minimize semantic heterogeneity, starting with characterizing its extent in a given data source. We identified the various sources of SH and found that characterizing SH provides a basis for understanding how to map and transform data to overcome these differences. Most of the DE in the HIV research domain were semantically heterogeneous, and most of the SH derived from semantically-equivalent/syntactically-different DEs. Expert-based solutions alone may be impractical, and data-driven methods may miss critical semantic relationships. We found that our data-driven method, SMASH, and expert-based methods are mutually beneficial for assessing SH among health-associated data, especially among semantically-equivalent/syntactically-different DEs. We recommend that health policy makers encourage the characterization of semantic heterogeneity in health related data, and we encourage the continued development of high-throughput expert-based and data-driven solutions, working in concert, that can address semantically-equivalent/syntactically-different DEs.

Contributions

William Brown III played a major role in all aspects of the research, conducted the analysis, devised the SMASH method, and was the primary author of the manuscript. Suzanne Bakken (PhD thesis sponsor) and Alex Carballo-Diéguez contributed to data dictionary identification, recruitment, acquisition and provision. Suzanne Bakken, David Vawdrey, and Chunhua Weng contributed to conceptualization and design of the overall research. Suzanne Bakken wrote the overall grant and secured funding for the project. All authors significantly contributed to the substantive review of the manuscript.

Acknowledgements

This work is supported by NIMH/NIDA R03- MH103957 (PI: Bakken). William Brown III is supported by National Library of Medicine (NLM) (T15-LM007079, PI: Hripcsak), and by a center grant from the NIMH to the HIV Center for Clinical and Behavioral Studies at New York State Psychiatric Institute and Columbia University [P30-MH43520; PIs: Ehrhardt (1987-2013) and Remien, PhD (2013-2018)]. The content is solely the responsibility of the authors and does not necessarily represent the official views of NLM, NIMH, or the NIH.

References

18. van der Loo M. A package for string distance calculation and approximate string matching (stringdist package in R) [Internet]. 2016 [cited 2016 Mar 10]. Available from: https://cran.r-project.org/web/packages/stringdist/stringdist.pdf
Affirming Proposed Variable Relationship Patterns in a Conceptual Nursing Model by Converting Qualitative Data to Causal Loop Diagrams

Jennifer A. Browne, Ph.D., RN-BC, CCRN

1University of Texas Health Science Center San Antonio, Texas

Abstract

Even with decades of use, there is minimal understanding about the impact that the use of Health Information Technology has on nursing work and workarounds. Reliance on quantitative methods has to some degree constrained our understanding by viewing phenomena from only one perspective. This multimethods research used qualitative data to develop causal loop diagrams and inform a Health Information Technology Workaround model. This approach can play an important role in generating an improved understanding of nursing clinical workflow and workarounds. This research strategy has not been identified in nursing literature to date, but perhaps will encourage future exploration and paradigm crossing. Investigating the use of causal loop diagrams and systems modelling in nursing can create an opportunity to enrich our insights and encourage scientific dialogue about the complexity of clinical workflow and the integration of Health Information Technology.

Introduction

The purpose of this paper is to describe a multimethod approach to model development and workflow diagramming. In informatics research there is a tendency to use a quantitative approach to investigations under the assumption that this is the way to reach significant conclusions. A flaw in this perspective is that it relies on linear, structured relationships to build knowledge about clinical care, a phenomenon that does not operate according to straight path algorithms. The study of complex adaptive systems (CAS) however is framed by dynamic interactions and complexity. It often addresses multiple levels and units of analysis; using multimethods to tell a story about how and why phenomenon evolved. Unfortunately, as a newer science, there are not many examples of CAS analysis in nursing.1

This paper describes the transformation of qualitative data into causal loop diagrams (CLD’s) in order to map the proposed variable relationships in a Health Information Technology Workaround (HITW) Model (See Figure 1) depicting health information technology (HIT) workarounds. This work was part of a larger dissertation study exploring HIT workarounds being used by nurses in intensive care. This study generated a multi-level conceptual framework through the specification of relevant variables and relationships essential for understanding what is actually happening with the use of HIT at the bedside. One aspect of analysis was the creation of causal loop diagrams (CLDs) in order to integrate feedback loops with the proposed HITW Model. This paper describes the development of these diagrams. The CLD’s are models that portray the behavior of variables in a system, presented as causal relationships and feedback loops.2 The premise of this work is that the feedback loops of a system should be understood in order to understand the system.3
The proposed HITW model is presented in Figure 1. This is a three level model with the micro level representing the patient, mezzo representing the nurse and macro the organization. This model is based on Arthur Stinchcombe’s constructing social theories work for functional theoretical explanations. The dependent variable in this work is the homeostatic variable (H). The variables (causes) that disrupt homeostasis are tensions (T) and the variable that tries to compensate for this and return the system to homeostasis is structure (S). Stinchcombe’s work allows the review of functional theory as an easily understood pattern of relationships between variables.

**Systems Thinking**

The complexities of healthcare can be overwhelming, yet in order to simplify workflow models we may be missing the real-world experiences of clinicians and the true complexity of care. For example, cosigning high risk medication should result in better patient safety, however as one nurse described, there may be unintended consequences such as altered or missing documentation: “I will not list all my IV medication titration changes that I made just to keep from having to beg someone to put in their password several times for each change”.

Quantitative approaches tend to break processes down, focusing on small numbers of linear relationships. Conversely, systems thinking looks at interactions and relationships between variables and expands our view of the ever-changing landscape of the phenomena. Systems are active, dynamic processes with multiple, variable interactions that will vary with time and environment. For instance, the same variable can act as an independent and dependent variable simultaneously in a systems thinking approach. Negative and positive feedback loops are a fundamental point of focus and can significantly influence systems behavior.

There is a formal process for systems thinking, systems dynamics, that blends modeling and simulation research with qualitative methods. This multimethod research is generally broken down into five parts. The first step is to identify variables from the qualitative data and secondly to develop a causal loop diagram (CLD) using those identified variables. From the CLD, stock and flow diagrams are constructed. The fourth stage is the development of the mathematical equations formulating the problem as designed in the stock and flow diagram. Finally, model simulation is performed using the predefined formulas. The scope of this work led only as far as to consider the presence and general implications of feedback loops with interactions between safe patient care, turbulence, workload, barriers, workarounds and HIT protocols. In order to explore relationships between variables, the methodology for the development of causal loop diagrams (steps one and two) from the coding of qualitative data was followed.
Methods

This study was conducted in collaboration with The American Association of Critical Care Nurses and approved by The University of Texas Health Science Center IRB. A sample of 307 Registered Nurses voluntary responded to an email survey consisting of two qualitative open ended questions followed by quantitative items measuring nurse characteristics, elements of nursing work, HIT problems, and patient safety. Multiple sources of data were used to compare, refine, and elaborate findings from both methods. The qualitative/ modeling techniques have been described in systems dynamics research over the last decade, yet they are still poorly specified. In this case the procedures described by Kapmeier and Yearworth and White were followed. As a strategy to help ensure rigor, the research outline offered by Kapmeier was followed. To enhance the validity of the multimethod research, legitimation process and checks guided each stage of the research.

The initial HITW model was designed using literature and personal experience. An internet survey was selected to collect data from the critical care nurses. The survey asked the nurse to describe, in their own words, the problem they experienced with HIT and the workaround used. Quantitative variables alone (from the literature) would have limited the data collected, and variables we knew to be important (such as workarounds) had no associated measures. It was clear that qualitative inquiry would be essential in modeling this complex system. The ability to follow up with participants was considered and rejected because assuring complete confidentiality was imperative to gaining truthful responses.

We were cautious in determining the a priori sample size because of concern over the amount of narrative transcripts we would receive. Nunally recommends sampling at least 10 times as many subjects as variables, and so the sample size was determined from quantitative methods and set at 290. Unlike true grounded theory, the qualitative analysis was conducted after the literature work, but before the quantitative analysis. Using a qualitative data analysis software, new behaviors or problems were identified and quantified. Themes and concepts were explored across texts to identify relationships. After reducing the volume of data during the coding process, categories were generated that could be linked to the concepts presented in the literature.

The work was divided into six exercises:

1. Qualitative data was coded from the survey transcripts. Specific qualitative variables (i.e., types of workarounds and frequency measures) were quantified and imported to IBM Corp. SPSS.
2. Reliability procedures were performed with each scale and exploratory factor analysis performed on the scales for turbulence and HIT problems. Inter-rater reliability was performed by different groups of experts to determine the internal validity of qualitative items, and to confirm agreement between quantitative items and qualitative descriptions. To confirm the agreement of preliminary workaround definitions, inter-rater reliability analysis was conducted by a panel of critical care nurses.
3. Quantitative correlation analysis of variable relationships and identification of relational patterns.
4. Qualitative code relations browsers were utilized to produce variable relationship matrices using MAXQDA software. The strength and direction of the relationships were informed by the narratives and quantitative analysis.
5. Development of multilevel causal loop diagrams representing the micro/ mezzo level and the mezzo/ macro level.

Results

The respondents were 87% female and 13% male. Fifty-eight percent of the nurses were 45 years old or greater. Almost 50% of the nurses had a bachelor’s degree in nursing. 20.6% an associate degree and 19.9% had a master’s degree. Nurse experience was midway between a proficient and expert. Intensive care specialties included adult, pediatric and neonatal. Patient acuity was reported as: 61.8% critical, 28.7% guarded and 9.2% stable. Workload of the nurse was reported as heavy (40%) and moderate (58%). There was a wide range of software represented including: KBMA (Allscripts), Carefusion, Cerner, Epic, Meditech, McKesson, Soarian, eICU, EndoTool and GlucoStabilizer.

Key variable definitions were confirmed using quantitative approaches; factor structures were identified and the reliability of new scales assessed. The key variables and associated survey items were: workload, turbulence, patient safety hazard, and HIT barriers/ problems. In previous work on the HITW model, the variable workload did not fully describe the amount and type of work nurses were performing. The variable turbulence was created to measure additional, unanticipated work that nurses perform. Examples of turbulence attributes include distractions,
interruptions, missing equipment and loss of information. The alpha coefficient for the 15 items was .751, suggesting acceptable internal consistency. Turbulence is connoted as the delta symbol in the HITW model. Four workaround variables developed during a previous pilot were confirmed and their specific attributes identified. These workarounds variables are: problem solving and intuitive workaround, and informal and formal communication.19 All proposed variable definitions and attributes were supported.

Once qualitative coding was completed, binary code matrices were produced (Table 1). These matrices allowed identification of relationship patterns and represented a starting point for investigation of mutual causality and feedback. This approach helped identify co-occurring concepts within and across groups. Table 1 identifies, for example, no relationships between patient safety risks, workarounds, and the lack of a unit secretary. There is also no relationship between the workload attributes of admissions, transfers and discharges and patient safety risk. The matrix did however identify relationships between turbulence (technology response time), patient safety risk and workarounds. Some interesting relationships were suggested. For instance, technology response time and ethical/moral conflict were the only two items associated with workarounds creating a patient safety hazard. On the other hand, the creation of a patient safety hazard caused by the primary problem was associated with 9 turbulence items such as loss of information, equipment issues and information overload.

Table 1. Binary code matrix; turbulence and workaround types

| Turbulence/ Resource Inadequate Training | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| Turbulence/Communication Technology | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Turbulance/Lack of Resources | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Turbulence/Lack of Information: Handoff | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Staff/Secretary | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Equipment & Supply Issue | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Turbulence/Information Overload | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Interpersonal Distractions | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Lack of Resources | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Lack of Information: Handoff | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Staff/Secretary | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Equipment & Supply Issue | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Turbulence/Information Overload | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Interpersonal Distractions | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Turbulence/Lack of Resources | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Turning to the quantitative items, correlations examined between workaround variables and demographic data (i.e., nurse characteristics) demonstrated no significant relationships. Quantitative variable analysis revealed that direct relationships exist between nursing workload, turbulence, the HIT barrier (problem) and patient safety hazards. For example, there were significant positive relationships between the HIT barrier (problem) and workload ($r = .32, N = 293, p = .000$) and HIT barrier and turbulence ($r = .33, N = 293, p = .000$). Turbulence was positively correlated with safety hazards ($r = .41, N = 293, p = .000$) and workload ($r = .48, N = 293, p = .000$). The weakest association was between workload and safety hazards ($r = .16, N = 29, p = .005$). There was overall agreement between the quantitative correlation findings and the qualitative binary matrix.

The correlational analysis supported a number of positive loops. An example of four positive reinforcing loops suggested by the data is presented in Figure 2. Positive reinforcing feedback loops are labeled as R in a CLD.7 As the number of HIT problems increase, so does workload. As turbulence increases, workload increases and as HIT problems increase, turbulence increases. Finally, as turbulence increases, risks of a patient safety hazard increases.

The analysis also supported a number of negative loops. Examples of four negative balancing loops are presented in Figure 2. These negative balancing feedback loops are labeled as B in a CLD.7 In these examples, as intuitive and informal communication workarounds increase, stress decreases. As workarounds increase, inefficiency decreases and as informal communication workarounds increase, complexity decreases. The correlation results provided direct and inverse relationship data to derive proposed feedback loops, however for
quantitative data to be useful and accurate a large amount of data is required. The qualitative data resulted in more representative and detailed matrices.

**Figure 2.** Positive (reinforcing) and negative (balancing) feedback loops

To begin modeling of the CLDs, qualitative matrices were produced for the micro and mezzo variables. Since the relationships can be in either direction, the nature of the relationships (direction) is determined by the researcher, but assessment of the strength of the category relationships can be determined by the number of times two categories are linked. In many cases, the quantitative correlational data supported the findings and suggested the direction of the relationship (direct or inverse). The more frequently the category is linked in the matrix result, then the more evidence there is that the two categories are related. The qualitative matrix example in Table 2 suggests, for example, that as the stress associated with HIT problems increase so does the patient safety risk (n=17). The strongest relationship in this table showed that as efficiency improved with the use of workarounds patient safety improved (n=38).

**Table 2. Qualitative Matrix Example;**

<table>
<thead>
<tr>
<th>WA= Workaround</th>
<th>Turbulence</th>
<th>WA decreases Stress</th>
<th>Problem Increases Stress</th>
<th>Problem reduce efficiency</th>
<th>WA reduce efficiency</th>
<th>WA improve efficiency</th>
<th>WA improve Safety</th>
<th>Problem Safety Risk</th>
<th>Added Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbulence Communication</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>21</td>
<td>12</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>WA decreases stress</td>
<td>16</td>
<td>0</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Problem increases stress</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>13</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>Problem reduce efficiency</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>WA reduces efficiency</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>WA improves efficiency</td>
<td>21</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>WA improve safety</td>
<td>12</td>
<td>1</td>
<td>13</td>
<td>9</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Problem safety risk</td>
<td>4</td>
<td>3</td>
<td>17</td>
<td>12</td>
<td>2</td>
<td>14</td>
<td>15</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Added Tasks/ Extra steps</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>12</td>
<td>18</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

A CLD of the micro and mezzo interfaces was developed based on the preliminary data produced from the qualitative matrix report and supported by the quantitative analysis (Figure 3). This CLD provides evidence that
protocols are interacting with barriers, workload and turbulence creating positive and negative feedback loops that also interact with safe patient care. This is by no means complete, but does give a sense of the complexity of the balancing and reinforcing loops and some of the primary relationships that were identified.

Figure 3. Causal loop diagram micro and mezzo levels

Some of the primary relationships identified in the micro/ mezzo causal loop diagram included:
1. Balancing loop 1 (B1): As turbulence increases, workarounds increase and problems decrease.
2. Balancing loop 2 (B2): As patient safety risks increases, workarounds increase. When workarounds increase, stress and inefficiency decrease and so safety risks decrease.
3. Balancing loop 3 (B3): As problems increase, workarounds increase. Workarounds act to improve the performance of HIT, decreasing process steps. Decreased process steps result in decreased problems.
4. Reinforcing Loop 1 (R1): As patient safety risks increase turbulence increases. Increased turbulence results in increased workload and that increases the patient safety risks.
5. Reinforcing Loop 2 (R2): As problems increase, turbulence increases. As turbulence increases so does workload. Increased turbulence and workload is associated with increases patient safety risk; increased risks contribute to an increase in problems.

Some of the primary relationships identified in the mezzo/ macro casual loop diagram (Fig.4) included:
1. Balancing loop 1(B1): As formal communication increases, process steps increase. As added steps increase there is a reduction in adherence to the protocols which forces an increase in formal communication.
2. Balancing loop 2 (B2): As protocol adherence increases so do the work process mismatches. These mismatches cause an increase in delays, increasing workarounds, and a reduction in adherence to protocols.
3. Balancing loop 3 (B3): As adherence to hospital protocols decrease, hospital reimbursement also decreases. When reimbursement decreases so does job security/ salary. As job security decreases, workarounds increase and are used not necessarily to increase accurate protocol use but to increase the perception that protocols are being followed.
4. Balancing loop 4 (B4): As care delivery delay increases, lost time increases. With an increase in lost time, there is an increase in workarounds. As workarounds increase the care delivery delays decrease.
5. Reinforcing loop 1 (R1): As workarounds increase, meeting the intent of the protocol increases. As compliance with protocols increase (or at least the appearance of compliance) then job security increases.
6. Reinforcing loop 2 (R2): As formal communication workarounds increase, added process steps also increase. As the steps increase so do the delays. When delays increase, the workarounds increase and this increases the use of formal communication workarounds.
7. Reinforcing loop 3 (R3): As workload increases, turbulence increases. An increase in turbulence causes an increased safety risk which then causes an increased errors. Increased errors cause an increase in workload.
Discussion

The use of qualitative modeling can provide an opportunity to consider situations that might not have been considered before. In quantitative inquiry the variables explored are pre-determined while causal loop diagramming allows for consideration of new behaviors during the study. For example, a number of nurses described that their year-end evaluations were directly tied to a compliance audit of scanning percentages and documentation. As this was explored, it was realized that nurses who could not perform the behavior were using workarounds to give the appearance that the protocol had been met.

The qualitative data supported the idea that nurses try to use HIT and the associated protocols to deliver safe patient care. It was also evident that nurses were attempting to comply with the HIT protocols even when the technology, added steps and work process mismatches made it difficult or impossible. When attempting to utilize Structure variables to achieve Homeostasis, nurses often encountered problems and turbulence which disrupted care delivery. In turn, nurses used workarounds as solutions for efficiency, complexity and time problems in an attempt to achieve Homeostatic outcomes.

Increased workload and reduced staffing were confirmed to add additional complexity, time pressure and inefficiencies further threatening safety and efficiency. Workarounds acted to mediate the relationships between turbulence, workload and patient safety. This was substantiated by balancing feedback loops between turbulence and workarounds and the time saved by use of a workaround in the absence of a problem. Current thinking is that patient safety risk increases when workarounds are utilized, but this analysis suggests that the opposite may also be true; workarounds are being used when a nurse recognizes a patient safety threat. It was also determined that workarounds were being used to comply with protocols. For example, it was reported that one protocol required that barcodes on all blood products be scanned prior to administration as part of the safety checks. When the blood bank had to split the blood product between 2 bags, there was no barcode on the second bag. The nurse entered another order for the same blood product in order to generate a barcode, and then canceled the order. Using this workaround, the nurse could follow the policy and scan the barcodes on both bags.

The CLD’s helped identify additional processes that the primary relationships might contribute to. For instance, formal communication workarounds are defined as written or oral communication that occurs through designated channels of the organization to address HIT systems barriers and disseminate protocol variations. Correlation analysis and CLD’s identified positive relationships between formal communication, extra process steps and care delivery delay. The CLD also identified that the administrative sanctioned workarounds (formal communication) indirectly reduced a nurses’ adherence to protocols. (Fig.4) This had not been apparent from the qualitative analysis alone. Finally, the development of CLD’s from code matrices also permitted us to see quite
rapidly where relationships did not exist. For instance, in nursing it is often assumed that inadequate HIT training is somehow associated with increased patient safety risk and workarounds. In fact, it is not uncommon to find mandatory education sessions following safety events. The code matrix (Table 1) did not identify a relationship between these variables.

Figure 5 presents an overlay of the CLD’s with the micro-mezzo HITW model. The feedback loops confirmed and clarified the relationships in the HITW model. The relationship between workload and adherence to HIT protocols was confirmed as bi-directional. Problems with S are filtering directly back to T (i.e., HIT performance, additional steps, staffing) further increasing workload and Tension. Turbulence interaction with model variables was apparent at the anticipated interfaces. The CLD’s supported the HITW model proposal that barriers (problems) and turbulence interfere with a nurses’ ability to utilize HIT to protect the patient from safety risk.

Figure 5. Causal loop diagrams and HITW model at the micro-mezzo level

One criticism of qualitative system dynamic modeling is the concern that wrong inferences might be attained. Causal loop structures can quickly become overly large and complex, masking the primary model behavior with excessive detail. It is recommended that a system archetypal model be maintained summarizing the essence of the model. In this case, the HITW model summarizes the relationships between HIT protocols, patient safety and workload and denotes the primary feedback loops derived from the quantitative and qualitative findings (Figure 5).

Implications

Qualitative inquiry allows identification and incorporation of variables that were perhaps never even considered at the onset of study. For example, documentation audits and the influence of nurses’ job evaluations on the use of workarounds was unexpected and not considered at the start of the study when quantitative variables were being designed. The qualitative analysis allowed the identification of this relationship during coding, inclusion in the matrix evaluations and timely integration of this interaction into the study.

Feedback loops occur when the output of a process becomes the input of another. Resulting behaviors, such as unintended consequences can be very hard to predict. The development of CLD’s assists in developing a view of the intertwining of problems and solutions and helps to anticipate the possible consequences of system or process use and changes. At the very least, interdisciplinary CLD development sparks a planning dialogue.

The CLD’s, although valuable, are only a first step. The next consideration must be the integration of time and time delay into the models. Integrating time into the analysis allows for understanding how much of a delay one variable can cause for the entire system. Matrices can be developed for time, and will assist in determining which variables are appropriate for intervention. This would be a most valuable asset for nursing and informatics planners.

Arising from the qualitative CLD development that displays links between cause and effect, are the stock and flow diagrams. The diagrams are more detailed and precise, relying on mathematical functions.
development of stock and flow diagrams. From stock and flow diagrams evolves the ability to run systems dynamic simulation and many simulation software products are evolving for precisely this type of use.

Although CLD’s are used as a stepping stone to more advanced analysis, there are some underlying truths and patterns that can be identified and utilized. CLD’s can be used to better predict consequences of change resulting from feedback loops. For example, CLD’s with a majority of negative feedback loops are more likely to display systemic resistance to changes or disturbances. Alternately, a system predominated by positive feedback loops can be highly unstable. Change should be approached cautiously and monitored for the emergence of new feedback loops. Different approaches to change management are recommended in each situation. Finally, as big-data and data analytics become more common place, the integration of system dynamics with data mining can more rapidly expand the knowledge discovery in databases.

Conclusions

Critical care nursing and healthcare in general is in a significant transformation. Rapid adoption of cutting edge technologies, new reimbursement systems and changes in the characteristics of nursing and physician jobs can stress an already complex healthcare system. The complexity of healthcare creates barriers to care delivery and, along with rapid change and poor workflow design, can inhibit a nurses’ ability to comply with the desired protocols envisioned by policymakers. When integrating HIT into clinical care and nursing workflow, a lack of theoretical underpinnings and a reliance on linear models has limited the ability to anticipate unintended consequences and consider the possible adaptations that nurses might make when using HIT.

This research offered one approach to integration of qualitative data into workflow analysis in order to explore variable relationships in a HITW model by creating causal loop diagrams. The foundational assumption in this work is that the nursing environments we are studying are dynamic complex adaptive systems that are constantly changing. The approach presented here offers one alternative to the structured linear methodology of nursing research by combining different methods in an attempt to better describe and understand the complexity of HIT use in critical care. The causal loop diagrams developed here, although elementary, do provide insights into nurses HIT workarounds in critical care. The research approach described in this paper is just the first step to more advanced analysis and the development of simulation models. Continuing to view our acute care environments as the complex adaptive systems that they are will require new approaches to research design and methods. This paper was one attempt to offer alternative ideas to the utilization of qualitative inquiry and discovery in nursing informatics research.
References

18. Browne, JA. The nature of turbulence and workload: conceptual and operational clarification. Summer Institute in Nursing Informatics, University of Maryland; 2016 July; Baltimore, Maryland.
Multi-Trajectory Models of Chronic Kidney Disease Progression

Philipp Burckhardt, MSc, Daniel S. Nagin, PhD, Rema Padman, PhD
Carnegie Mellon University, Pittsburgh, PA

Abstract

An ever increasing number of people are affected by chronic kidney disease (CKD). A better understanding of the progression of CKD and its complications is needed to address what is becoming a major burden for health-care systems worldwide. Utilizing a rich data set consisting of the Electronic Health Records (EHRs) of more than 33,000 patients from a leading community nephrology practice in Western Pennsylvania, we applied group-based trajectory modeling (GBTM) in order to detect patient risk groups and uncover typical progressions of CKD and related comorbidities and complications. We have found distinct risk groups with differing trajectories and are able to classify new patients into these groups with high accuracy (up to $\approx 90\%$). Our results suggest that multitrajectory modeling via GBTM can shed light on the developmental course of CKD and the interactions between related complications.

1 Introduction

Chronic Kidney Disease (CKD) is a growing burden for the national health-care sector. Today, it is estimated that more than 11% of the US adult population have some degree of CKD, and projections indicate that more than 50% of those aged 30 to 64 years will develop CKD. With costs amounting to $49.2 billion in the United States in the year 2011 for the treatment of End Stage Renal Disease (ESRD), the final stage of CKD, it is paramount to gain deeper insights into the progression of the disease in order to facilitate the development of new preventive care approaches.

The increasing adoption of Electronic Health Record Systems (EHRs) in recent years, fueled by the promise of cost savings, increased efficiency and better communication between the various healthcare providers, has resulted in the accumulation of a massive amount of structured data on patients and their disease progressions. It has been argued that the identification and the quality of care of CKD patients could be improved by an effective utilization of EHRs.

Using a rich data set consisting of the EHRs for more than 33,000 patients from a leading community nephrology practice in Western Pennsylvania, this is the first study that applies group-based trajectory modeling (GBTM) in order to uncover typical progressions of CKD and related comorbidities and detect patient risk groups. By modeling biomarkers not only for CKD, but also complications typically linked to it, we aim to obtain a fuller, multi-dimensional picture of its progression. Specifically, we use the estimated glomerular filtration rate (eGFR) as a biomarker for CKD and other appropriate laboratory-based biomarkers for its complications.

Historically, CKD progression was assessed via patient's serum creatinine levels. However, serum creatinine is not a good measure of kidney function, so laboratories nowadays report eGFR in addition. eGFR can be calculated from the level of serum creatinine and patient characteristics such as age, gender and race. This change in reporting was associated with an increase in first nephrologist visits, but eGFR by itself is not sufficient for guiding decision-making on the care of kidney patients.

A multi-dimensional approach is required since CKD patients tend to have numerous comorbidities. In fact, a large fraction of patients suffering from CKD do not progress to the later stages of the disease but die prematurely due to these comorbidities and complications. Thus, any treatment should take into account the contemporaneous progression of CKD and its varying complications at differing levels of severity that patients experience. This poses a major challenge, though, since care coordination amongst various medical professionals is involved. Discerning disease patterns via interpretable and accessible statistical models has the potential to alleviate the communication challenges between these stakeholders. To facilitate this, we develop and test a statistical model which can be used to track not only the progression of CKD but also the development of several complications. This can be accomplished by monitoring the levels of biomarkers for the considered complications.

In clinical research, group-based trajectory models (GBTM) are increasingly used to model the development of a clinically important indicator over time, with the goal of identifying groups of individuals sharing a common trajectory.

1737
Originally devised by Nagin and colleagues as a criminological tool for classifying criminal careers, this approach lends itself towards usage in a wide range of disciplines, including clinical research.\textsuperscript{8} At their core, GBTMs are an example of a class of statistical models called finite mixture models. More precisely, GBTMs are mixtures of regression models applied to longitudinal data, where the likelihood of a time series for an individual is assumed to be a mixture of linear models that include both an explicit time variable as well as other covariates. This way of analyzing biomarker trajectories is different from other approaches such as latent growth curve modeling, which rests upon the assumption of a common functional form for the trajectories for all individuals, with its parameters varying randomly. Growth curve models, which can be formulated either as structural equation models or as multilevel models, effectively assume all individuals to belong to a single class. In contrast, GBTM have a finite number of classes, each of which is characterized by its own parametric trajectory. Given that our interest lies not in the study of the variability between subjects, but in discovering the typical marker trajectories that we suspect to exist, GBTMs are more appropriate for our needs.

The data set spans four years of patient data and includes diagnoses, lab results and patient characteristics. After restricting the data set to all patients who were diagnosed with CKD Stage III during the covered time span, we removed all transplant patients as their trajectories differ significantly from the rest of the patients. The final cohort consists of 1,944 individuals.

We fit a single, joint multi-group trajectory-model for five biomarker time series, and use an appropriate model selection procedure to identify a parsimonious and interpretable model. Covariates such as patient characteristics (age, weight, gender, race, etc.) and binary indicators on whether the patient suffers from the comorbidities of diabetes and hypertension are included in addition to the biomarker time series. By fitting a joint model for all markers, we build upon earlier work on single and dual trajectories, whose results hinted at the usefulness of GBTMs for risk stratification of CKD patient populations, even though only one or two markers were used at a time.\textsuperscript{9} The constructed trajectory model is necessarily a simplification of the real world, but ideally one which captures the main distinctive trajectories of disease progression one would typically encounter in patients. Its use could enable better screening of patient populations for high-risk individuals and lead to insights about the interplay between the various complications that are suffered by CKD patients.

\section{Data}

We are working with a rich clinical data set of patients from a leading nephrology practice in southwestern Pennsylvania. The total number of unique patient records is 33,882. The data set contains information about patient characteristics (age, gender, weight and height) as well as all of their lab measurements with associated dates.

The lab results span a total of 18 quarters, ranging from the years 2009 to 2013. The patient population is split almost evenly among female and male patients. Most patients are retirees, with median age of 70 years. 94\% of the patients are white. About half the patients have a diagnosis of CKD Stage III while the remaining are in more advanced stages of CKD.

For our analysis, patients diagnosed with CKD Stage III between January 1, 2009 and November 19, 2012 were selected on the grounds that this diagnosis usually marks the point when patients start showing first disease symptoms and are referred to a nephrologist. Patients in CKD Stages I and II show almost normal kidney function, and are therefore rarely diagnosed. A doctor will only make these diagnoses by linking additional evidence like proteinuria or haematuria, which are not sufficient indicators of CKD.

Patients who received a kidney transplant on or after January 1, 2009 were removed from the analysis since their level of kidney function differs sharply from those who did not receive a transplant and keeping them will likely distort the identified disease progressions.

Historically, serum creatinine levels have been used as a marker for CKD: As kidney function deteriorates, blood levels of creatinine typically rise. Tracking the estimated Glomerular Filtration Rate (eGFR) has replaced the creatinine test as a more reliable means to detect early kidney damage and is a method suggested by the National Kidney Foundation (NKF). eGFR can be derived from the creatinine level and additional variables such as age, sex and race and possibly height and weight. For this study, eGFR was calculated from serum creatinine levels using the CKD-EPI equation.
The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) developed and validated this equation, which is more accurate than the previously used Modification of Diet in Renal Disease (MDRD) Study equation, although it is not clear whether it improves risk prediction.11

Creatinine values larger than 10 mg/dl were removed because these values were likely entered incorrectly. We only keep measurements collected on or after the date the patient was diagnosed with CKD Stage III. Typical complications of CKD include Anemia,12 Secondary Hyperparathyroidism,13 Hyperphosphatemia14 and Metabolic Acidosis.15 Our data set allows us to track each of these via corresponding lab measurements. The respective lab measurements of Hemoglobin (HGB), parathyroid hormone (PTH), phosphate (PO4) and carbon dioxide (CO2) can be used as markers for the considered complications. In the cohort, 1,367 patients develop Anemia, while 1,476 of them are diagnosed with Secondary Hyperparathyroidism. 438 patients have Acidosis, but only a hundred suffer from Hyperphosphatemia. Table 1: Averages for the five considered biomarkers across time periods (quarterly). Standard deviations are displayed in parentheses. The overall data availability is displayed in the second row.

<table>
<thead>
<tr>
<th>Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>eGFR</td>
<td>37.87</td>
<td>38.33</td>
<td>38.64</td>
<td>38.55</td>
<td>38.70</td>
<td>38.42</td>
<td>38.54</td>
<td>38.47</td>
<td>38.29</td>
<td>38.18</td>
<td>38.13</td>
<td>37.84</td>
<td>37.64</td>
<td>36.91</td>
<td>36.92</td>
<td>35.95</td>
<td>35.29</td>
<td>33.71</td>
</tr>
<tr>
<td>PTH</td>
<td>(2.59)</td>
<td>(1.98)</td>
<td>(1.67)</td>
<td>(1.58)</td>
<td>(1.84)</td>
<td>(1.69)</td>
<td>(1.65)</td>
<td>(1.66)</td>
<td>(1.78)</td>
<td>(1.66)</td>
<td>(1.67)</td>
<td>(1.77)</td>
<td>(1.68)</td>
<td>(1.66)</td>
<td>(1.66)</td>
<td>(1.70)</td>
<td>(1.75)</td>
<td>(1.83)</td>
</tr>
<tr>
<td>PO4</td>
<td>36.34</td>
<td>36.26</td>
<td>36.15</td>
<td>36.23</td>
<td>36.20</td>
<td>36.14</td>
<td>36.19</td>
<td>36.23</td>
<td>36.30</td>
<td>36.34</td>
<td>36.31</td>
<td>36.31</td>
<td>36.29</td>
<td>36.27</td>
<td>36.06</td>
<td>36.26</td>
<td>36.18</td>
<td>36.37</td>
</tr>
<tr>
<td>Availability</td>
<td>0.267</td>
<td>0.43</td>
<td>0.546</td>
<td>0.45</td>
<td>0.529</td>
<td>0.409</td>
<td>0.412</td>
<td>0.338</td>
<td>0.405</td>
<td>0.292</td>
<td>0.289</td>
<td>0.255</td>
<td>0.287</td>
<td>0.211</td>
<td>0.196</td>
<td>0.149</td>
<td>0.13</td>
<td>0.035</td>
</tr>
</tbody>
</table>

All outcome variables were averaged by quarter to deal with the relative data scarcity and infrequency of the lab measurements. In addition, this procedure yields measurements on a discrete time scale as required by the group-based trajectory model. For each biomarker, Table 1 displays the average value across time periods and its standard deviation.

Patients who do not have at least four observations in the considered time range were removed. Even for the remaining patients, observations are scarce: On average, for each patient we have only measurements for 34 quarters whenever there was a missing quarter between them. Doing so increased the percentage of available data from 42% to 62.42%.

After data cleaning and processing, 1944 of the patients diagnosed with CKD Stage III on or after January 1, 2009 were used for model fitting.

3 Methods

Single Trajectory Model

Let \( Y_j \) define the multivariate response variable for the \( j \)-th marker, for example eGFR. Then, \( Y_j = (Y_{j,1}, \ldots, Y_{j,18})^T \) is a vector of length 18, which holds the quarterly lab results for the marker in question. For a single response, the group-based trajectory model posited by Nagin16 assumes the following density for a sequence of longitudinal measurements \( y_j = (y_{j,1}, \ldots, y_{j,18}) \):\[
f(y_j) = \sum_{k=1}^{K} p_k f_{Y_j}(y_j | C = k), \tag{1}
\]
where \( K \) denotes the total number of groups, \( p_k \) the probability of belonging to group \( k \) and \( f_{Y_j}(y_i | C = k) \) the conditional density of the observed data given class \( k \).

The probabilities \( p_k \) of this mixture model are not estimated directly, but related via the softmax function to a \( K \)-
dimensional vector $\theta$ of class coefficients and time-stable covariates $x$ with associated weight vectors $w_k$:

$$p_j = \frac{e^{\theta_j + x^Tw_j}}{\sum_{k=1}^{K} e^{\theta_k + x^Tw_k}}.$$  

(2)

The response vector for each outcome is modelled as a multivariate normal random variable

$$Y_j | C = k \sim \mathcal{N} \left( \mu_{jk}, \sigma_j^2 I \right),$$  

(3)

where the elements of the mean vector are related to the period $t$ (= time in quarters since diagnosis of CKD Stage III) of the individual patient as follows:

$$\mu_{jkt} = \beta_{jk0} + \beta_{jk1} t + \beta_{jk2} t^2 + \beta_{jk3} t^3.$$  

(4)

As can be seen, the group-based trajectory model is based on the assumption that the trajectories in each group have a simple polynomial form. From our experience, a polynomial order above three is rarely necessary, which is why we have constrained the model to have cubic terms at most.

### Multi-Trajectory Model

This is one of the first works in which the group-based trajectory model devised by Nagin\textsuperscript{16} is used not to model just a single time series, but jointly the trajectories of multiple outcomes.\textsuperscript{17} In this model extension, the density for $J$ outcomes becomes

$$f(y) = f(y_{j=1},...,J) = \sum_{k=1}^{K} p_k \prod_{j=1}^{J} f_{Y_j}(y_j | C = k).$$  

(5)

Model fitting and inference is carried out via the `traj` procedure from the Stata package of the same name,\textsuperscript{18} which implements a Newton-Raphson optimization algorithm for maximum likelihood estimation. For $N$ observations $y^{(1)}, \ldots, y^{(N)}$ of all outcomes, the maximum-likelihood optimization problem is

$$\max_{\theta, \beta, \sigma, \omega} \mathcal{L} = \max_{\theta, \beta, \sigma, \omega} \prod_{i=1}^{N} f_{y^{(i)}}(y^{(i)}; \theta, \beta, \sigma, \omega).$$  

(6)

Following the suggestion given by Jones et al., the Bayesian information criterion (BIC) is used to perform model selection and to determine the number of groups $K$.\textsuperscript{18}

### Model Predictions

The aforementioned procedure yields a multi-response model, which fits trajectories for all five biomarkers and incorporates time-stable covariates, namely demographic variables and indicators for the existence of diabetes and hypertension. New patients can be classified using the posterior probabilities of group membership, which can be computed by using Bayes’ rule as

$$\Pr \left( C = l \mid \{Y_j \mid j=1,\ldots,J \} \right) = \frac{p_l \prod_{j=1}^{J} f_{Y_j}(y_j \mid C = l)}{\sum_{k=1}^{K} p_k \prod_{j=1}^{J} f_{Y_j}(y_j \mid C = k)},$$  

(7)

where the number of outcomes $J$ is equal to five in our case. For the conditional densities, it follows from Equation (3) that

$$f_{Y_j}(y_j \mid C = k) = \prod_{t=1}^{T} \phi \left( \frac{y_{jt} - \mu_{jkt}}{\sigma_j} \right),$$  

(8)

in which $\phi$ is the density function of the standard normal distribution. For the data used for model fitting, $T = 18$. However, in order to form predictions, we cannot calculate the group membership conditional on $Y_i$ for all time
periods as they might not be available yet. By conditioning only on measurements up to the observed time, we can calculate posterior probabilities that take all currently available information into account.

Similarly, if for any outcome the time series $Y_t$ contains missing values or starts at a later time, the product will only be taken over the observed data. This allows us to generate predictions for new data points for which we have not yet observed the trajectories for 18 time periods, which is important given the goal of detecting high-risk patients early on and not when it may already be too late.

4 Results

![Trajectories for eGFR](image)

The estimated trajectories for eGFR from the multi-trajectory model are displayed in Figure 1. While all patients in our cohort have been diagnosed with CKD Stage III and thus suffer from kidney damage, it becomes clear that some patients belong to groups characterized by trajectories that show almost no change in eGFR values (groups 5-8), whereas the kidney function of patients in other groups deteriorates significantly after they were first diagnosed (groups 1-4).

The number of groups was determined using Bayesian information criterion (BIC) as the model selection criterion, as was the order of the polynomial terms and the inclusion of time-stable covariates. Using BIC in the model search is supposed to help build a sparse but at the same time sufficiently large model to accurately reflect the data at hand. Starting from a full model with cubic polynomials and all available covariates, variables were removed from the model if they were statistically insignificant and their exclusion improved the BIC score. Eight groups were selected in total, and of the available time-stable covariates, gender, BMI, age, and indicators for being black, being diagnosed with hypertension and having diabetes all ended up in the model. First-order interaction terms between the indicators were not significant and hence we dropped them from the model.

Table 2: Displayed are the group sizes as a proportion of the entire population in the second column of the table, as well as the averages of the included time-stable covariates inside of each group in columns three to eight.

<table>
<thead>
<tr>
<th>Group</th>
<th>Proportion</th>
<th>Females</th>
<th>Black</th>
<th>Mean Age</th>
<th>Mean BMI</th>
<th>Hypertension</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10</td>
<td>0.52</td>
<td>0.09</td>
<td>68.71</td>
<td>31.50</td>
<td>0.99</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>0.12</td>
<td>0.69</td>
<td>0.03</td>
<td>73.96</td>
<td>30.98</td>
<td>0.99</td>
<td>0.53</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>0.46</td>
<td>0.04</td>
<td>71.09</td>
<td>30.04</td>
<td>0.96</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>0.13</td>
<td>0.60</td>
<td>0.07</td>
<td>76.44</td>
<td>32.16</td>
<td>0.98</td>
<td>0.52</td>
</tr>
<tr>
<td>5</td>
<td>0.16</td>
<td>0.31</td>
<td>0.02</td>
<td>70.89</td>
<td>31.64</td>
<td>0.96</td>
<td>0.43</td>
</tr>
<tr>
<td>6</td>
<td>0.09</td>
<td>0.46</td>
<td>0.02</td>
<td>68.49</td>
<td>31.69</td>
<td>0.98</td>
<td>0.56</td>
</tr>
<tr>
<td>7</td>
<td>0.13</td>
<td>0.50</td>
<td>0.11</td>
<td>73.27</td>
<td>30.97</td>
<td>0.98</td>
<td>0.39</td>
</tr>
<tr>
<td>8</td>
<td>0.12</td>
<td>0.18</td>
<td>0.05</td>
<td>67.27</td>
<td>31.27</td>
<td>0.93</td>
<td>0.38</td>
</tr>
<tr>
<td>Overall</td>
<td>1.00</td>
<td>0.46</td>
<td>0.05</td>
<td>71.45</td>
<td>31.26</td>
<td>0.97</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table 2 shows that the groups are roughly equal in size, but differ with respect to some demographic variables. The
differences are most pronounced for the racial makeup, with Blacks being over-represented in groups one and seven. Even though the differences might be more nuanced for some of the other variables, the null hypothesis that there are no differences among the groups is rejected for each variable at a significance level of 5% when conducting an ANOVA. As one can also see from Table 2, patients with diabetes are more likely to belong to the high-risk groups compared to groups such as seven and eight, which are characterized by stable and better eGFR values.

![Fitted Trajectories of the Eight-Group Multi-Trajectory Model](image)

**Figure 2:** Fitted trajectories of the eight-group multi-trajectory model for the five considered biomarkers

Since patient characteristics seem to tell only part of the story, inspecting the estimated trajectories for the biomarkers might reveal deeper insights into the risk factors associated with CKD. All estimated trajectories are displayed together in Figure 2. Background shading indicates a ranking of the group trajectories in terms of whether marker values are better or worse than those for the other groups. As emphasized by the color coding, groups one to four have worse values compared to the rest for most markers, with group one being uniformly worse and the other three each having one or more markers for which values are better, indicating that patients might not (yet) be afflicted by the corresponding complication.

For example, groups one, two and four are all characterized by patients suffering from Anemia, as evidenced by their low hemoglobin values. It has been argued that there are “potentially severe consequences of anemia in CKD”. Indeed, these three groups develop worse in terms of eGFR than the rest, except for group three. Yet, the remaining markers also show signs for the complications, which furthermore illustrates that patients suffer not only from CKD, but often a variety of other chronic conditions.

The complications are associated with how fast CKD itself progresses: For example, high-risk group one shows the worst trajectories for all considered markers, and a majority of its members do progress to at least CKD Stage IV, if not ESRD, as can be glanced from Table 3.

As expected, groups with worse trajectories for the considered markers (such as groups one to four) have a larger percentage of patients ending up in a more serious disease stage.

For each diagnosis displayed in Table 3, a chi-squared test rejects the null hypothesis of independence between the variable in question and group membership ($p < 0.001$). Knowing the risk group a patient most likely belongs to is therefore inextricably tied to observable outcomes and diagnoses.
Table 3: Contingency table for group membership and the final CKD diagnosis of the patient. A patient can experience either no progression of his CKD status or advance to either Stage IV or Stage V, which is also called End-Stage Renal Disease (ESRD).

<table>
<thead>
<tr>
<th>Group</th>
<th>No Decline</th>
<th>Stage IV</th>
<th>ESRD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>106</td>
<td>73</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>8.2%</td>
<td>54.4%</td>
<td>37.4%</td>
<td>10.0%</td>
</tr>
<tr>
<td>2</td>
<td>118</td>
<td>97</td>
<td>11</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>52.2%</td>
<td>42.9%</td>
<td>4.9%</td>
<td>11.6%</td>
</tr>
<tr>
<td>3</td>
<td>131</td>
<td>138</td>
<td>15</td>
<td>284</td>
</tr>
<tr>
<td></td>
<td>46.1%</td>
<td>48.6%</td>
<td>5.3%</td>
<td>14.6%</td>
</tr>
<tr>
<td>4</td>
<td>111</td>
<td>131</td>
<td>16</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>43.0%</td>
<td>50.8%</td>
<td>6.2%</td>
<td>13.3%</td>
</tr>
<tr>
<td>5</td>
<td>246</td>
<td>64</td>
<td>3</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td>78.6%</td>
<td>20.4%</td>
<td>1.0%</td>
<td>16.1%</td>
</tr>
<tr>
<td>6</td>
<td>140</td>
<td>37</td>
<td>1</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>78.7%</td>
<td>20.8%</td>
<td>0.6%</td>
<td>9.2%</td>
</tr>
<tr>
<td>7</td>
<td>244</td>
<td>9</td>
<td>3</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>95.3%</td>
<td>3.5%</td>
<td>1.2%</td>
<td>13.2%</td>
</tr>
<tr>
<td>8</td>
<td>234</td>
<td>0</td>
<td>0</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>100.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>12.0%</td>
</tr>
<tr>
<td>Total</td>
<td>1240</td>
<td>582</td>
<td>122</td>
<td>1944</td>
</tr>
</tbody>
</table>

One promising use of trajectory models for health outcomes is to detect sub-populations of patients who are particularly at risk and might benefit from early interventions. In Figure 3, the changes in misclassification rate are displayed when not all time periods of the data are used, but only those up to a chosen period. As expected, the misclassification error for the full training data approaches zero when all periods of data are used, which is tautological because the group memberships are determined by assigning each patient to the group with the highest posterior probability. However, when we use fewer and fewer data points, the misclassification rate goes up since the patients cannot be placed definitely in one of the groups yet. The negative slopes of the two plotted lines are large in magnitude, though, indicating that the model does dramatically better as more data becomes available. To demonstrate that this pattern occurs also on previously unseen data, we set aside 20% of the observations as a test set and refit the model on the remaining 80% of the data. The misclassification rate on the test set is displayed as the solid line in the plot. It shows the same curvature, with the main difference being that the error does not approach zero but instead converges to 10% after all periods are taken into account. The group memberships deemed as ground truth are obtained from fitting a multi-trajectory model on the entire data set. These results are encouraging evidence that the model has useful predictive power.

Figure 3: Model performance of the multi-group trajectory model as a function of the number of time periods.

Given our interest in obtaining predictions for individual patients, it is convenient that the laws of probability lead
naturally to individualized posterior probabilities of group membership for a patient given his or her lab measurements and demographic variables, which we can track over time. In Figure 4, posterior probabilities are displayed for an average patient with the values for the biomarkers set to the overalls means of the entire cohort. The average patient has an age of 71.45 years and a BMI of 31.26. All other covariates are set to their baseline levels, i.e. the patient is male, has neither diabetes nor hypertension and is white. Starting with the second quarter, the posterior probability of belonging to group three exceeds all others. The final group label emerges as the top choice after only two time periods, suggesting that the trajectory model quickly converges to assign a patient to one of the groups as new data comes in. This is an encouraging observation, as it shows that trajectory modeling might be useful not only in analyzing outcomes ex post, but also beforehand.

Figure 4: Posterior probabilities as a function of the number of time periods for the average patient in our patient population.

Profiling the average patient might not be of much interest in itself, but is illustrative of the potential uses of the model in predicting group membership for individual patients and sub-populations that are of special concern.

5 Conclusion

Using group-based trajectory modeling (GBTM) to detect patient risk groups for CKD sharing comparable trajectories for the estimated glomerular filtration rate (eGFR) and markers for four complications of CKD, we have identified eight groups in a multi-trajectory model for the five outcomes. The results are consistent with the previous study by Padman et al., but build on those findings with the application of the multi-trajectory to the five jointly considered outcome variables, thus offering a more detailed perspective. The eight groups are characterized by distinct trajectories and patient demographics. For the high-risk groups 1-4, diabetes is prevalent, whereas it occurs less often for patients in the lower-risk groups 5-8. Interestingly, the proportion of black patients is largest in the two most extreme groups, one and eight, which are characterized by both the best and the worst development of the tracked biomarkers.

It is a well known fact that Blacks are more prone to develop CKD than non-Blacks. Yet, at the same time it has been observed that Blacks have a survival advantage in the ESRD population, the reasons of which are not entirely clear. Against this background, one might have expected the larger proportion of Blacks in the high-risk group, but not necessarily in group seven.

In an analysis of the progression of renal failure, Hannedouche et al. found gender to be correlated with disease progression, with male patients progressing significantly faster. We did find some large differences in the gender makeup of the eight groups, with group eight standing out in particular: Only 18% of patients assigned to this group were females. While patients in this group are younger than in the others, in line with the finding that males develop CKD faster, their eGFR rate is surprisingly stable and only few of its members progress to stage 4 or stage 5 of the disease. This is a somewhat unexpected finding, which could be explained by the fact that groups seven and eight might be separated just because of the difference in hemoglobin levels between the sexes.

It is known that CKD prevalence increases drastically with age and it has been speculated that elderly people might be more susceptible to CKD. Yet, Hallan et al. found low eGFR to be “independently associated with mortality and
ESRD regardless of age across a wide range of populations” in patient cohorts selected for CKD.\textsuperscript{21} This is consistent with our results, which do not enable us to use age to discriminate between the high-risk and low-risk groups.

All in all, it seems that the value of demographic variables for risk stratification is not high, and that more emphasis should be placed on the trajectories of the different biomarkers. At the same time, given the complex nature of CKD, additional biomarkers such as Urine-Albumin Creatinine Ratio (UACR) may need to be included in future studies.

We have shown that patients can be classified into the detected groups with an accuracy of more than 50\% after a year of data. Collapsing the groups into the high-risk and low-risk communities of groups 1-4 and 5-8, this number increases to approximately 82\%. For previously unseen patients, accuracy increased to 90\% after using all time periods as evaluated against a gold standard data set obtained by fitting the GBTM on all observations. These are encouraging signs, which indicate that GBTM could help with population risk stratification as well as to assess individual patient’s risk. For example, our model could be used for predicting stage progression (see the work by Perotte et al.\textsuperscript{22}) or another outcome like the need for a transplant. To avoid distorting the results, this latter case would require the inclusion of transplant patients into the cohort, which could be achieved by incorporating all their lab values until the day of their transplant. Here a strength of GBTMs is rendered visible: The ability to handle missing data.

One limitation of this study is a lack of mortality data for the patient population, which might have caused a bias in our results. Also, results might not generalize to other populations, given that the patients in our data set are fairly homogeneous in terms of demographic variables such as race and age. To establish that the identified patterns are externally valid, analyses based on samples with different demographics should be conducted. In this study, we had to deal with several data quality issues of the EHR such as erroneous patient identifiers and wrongly encoded lab values. Since it is not possible to rule out that some issues might have persisted, further analyses on similar data sets should be undertaken to replicate the results.

GBTMs could be extended in several ways, too. Assuming a polynomial trajectory is may not reflect reality well, but is a convenient assumption since it makes model fitting tractable and improves interpretability of the model. Alternatively, one could consider a non-parametric approach, at the expense of interpretability. In addition, the model assumes that conditional on group membership, the elements of $Y_j$ are uncorrelated with each other. Weakening this assumption and permitting the off-diagonal elements of covariance matrix $\sigma_j^2 I$ to be non-zero might yield a more realistic picture of the true underlying data generating process. Yet, the assumption is not as restrictive as it might appear on first glance: Outcomes are not modelled to be conditionally independent at the population level. The model merely assumes that conditional on the latent group membership, the Gaussian noise added to the trend line stays constant over time.

The analysis could also be enhanced by exploiting the rich data available inside the EHRs of each patient, most of which is unstructured. For example, clinical narratives written after each patient’s visit might reveal insights which the biomarkers themselves would not give away. Alternatively, one could investigate whether medications are associated with a decline or increase in one of the trajectories. This was not done in the present study because of significant difficulties posed by the high-dimensionality of the problem, which would require dimensionality reduction by mapping individual drug names to drug classes. This is the subject of a follow-up study conducted at the Heinz College at CMU.

While group-based trajectory modeling (GBTM) has seen increasing adoption in clinical research, to the best of our knowledge, this study is the first to jointly model multiple outcomes, thereby providing a fuller picture on patient’s disease progression. Given our results, we believe that multi-trajectory models provide a simple but powerful tool for risk stratification of patients as they allow identification of a finite number of groups with distinct trajectories for the developmental course of a disease and relevant predictors of group membership. Hence, GBTMs provide interpretable summaries of typical disease progressions and might help in the development of targeted, proactive interventions for patients in all risk groups.

**Acknowledgement**

We would like to thank the physicians and staff from Teredesai, McCann & Associates for providing the data for this study and their valuable input and suggestions on our analysis.
This study was designated as Exempt by the Institutional Review Board at Carnegie Mellon University.

References

PREMIX: PRivacy-preserving EstiMation of Individual admixture

Feng Chen, Ph.D.∗, Michelle Dow†, Sijie Ding‡, Yao Lu§, Xiaqian Jiang, Ph.D.†, Hua Tang, Ph.D.¶, Shuang Wang, Ph.D.†

1Department of Biomedical Informatics, UC San Diego, La Jolla, CA
2Department of Electrical and Computer Engineering, UC San Diego, La Jolla, CA
3Department of Genetics, Stanford University, Stanford, CA

Abstract

In this paper we proposed a framework: PRivacy-preserving EstiMation of Individual admixtiture (PREMIX) using Intel software guard extensions (SGX). SGX is a suite of software and hardware architectures to enable efficient and secure computation over confidential data. PREMIX enables multiple sites to securely collaborate on estimating individual admixture within a secure enclave inside Intel SGX. We implemented a feature selection module to identify most discriminative Single Nucleotide Polymorphism (SNP) based on informativeness and an Expectation Maximization (EM)-based Maximum Likelihood estimator to identify the individual admixture. Experimental results based on both simulation and 1000 genome data demonstrated the efficiency and accuracy of the proposed framework. PREMIX ensures a high level of security as all operations on sensitive genomic data are conducted within a secure enclave using SGX.

Introduction

Identifying the demographic histories of patients is an important problem arising in biomedical research. For example, given the accurate ethnicity information, researchers can better understand whether certain populations are more susceptible to particular disease or most likely to benefit from certain therapeutic interventions1. Understanding the individual admixture from different ancestries is also important for researchers who conduct case-control association studies2. Electronic medical records (EMRs) can provide clinicians and biomedical researchers a new perspective in studying associations with the symptom or medication use. However, research studies based on the races/ethnicity from EMRs often faces problem with missing or inaccurate self-described information1. Hispanics, for example, represent an admixed group between Native American, Caucasian and African. In addition, African-Americans represent another large admixed group. Researchers have shown that the individuals within the Hispanics or African-Americans groups did not form a distinct subgroup, but clustered variously within the other groups4. As a result, the self-report ethnicity information in EMRs may not provide the most accurate characterization of patients.

Genome-wide association studies (GWAS) provide a powerful tool for identifying genetic biomarkers which reflects an individual’s ethnicity by applying admixture models on allele frequencies of SNPs5–2. A basic assumption for ethnicity testing is that any current individual genome or population is a mixture of ancestries from past populations7. Population methods developed according to the amount of loci that can be traced back to a certain ancestry population is largely used. Companies such as 23andMe8 or Ancestry DNA9 have been the major autosomal DNA tests existed to reveal the ancestry of an individual. However, it is usually infeasible for researchers to scan for every patient’s ethnicity through these expensive tests. Rapid advances in sequencing technologies enable the meaningful use of human genomic data in a wide range of healthcare and biomedicine applications. Reuse existing genomic data of patients to identify patient ethnicity or improve the accuracy of self-report information can significantly improve the data quality in research study that requires population stratification.

The research team of 23andMe published 22 population-specific common SNPs that can reflect demographic histories5. The study was done from the self-reporting, participant-driven data gathered on the Web, and associations were discovered for the hair color, eye color, and freckling (in the genes OCA2, HERC2, SLC45A2, SLC24A4, IRF4, TYR, TYRP1, ASIP, and MC1R)5. Similar researches with SNPs associations are done by Yaeger et al.10 and Kosoy et al.11, which found 107 and 128 SNP race/ethnicity-related biomarkers, respectively. For example, Yaeger et al.10 investigated with 50 African Americans and 40 Nigerians as their subjects. Ancestry informative markers (AIMs) used in their study were based on bi-allelic SNPs that were selected from the Affymetrix 100K SNP chip based on “informativeness” of ancestry’s genotype data. Informativeness12 between multiple population groups

∗ Both authors share the first authorship
was determined using mutual information. Furthermore, Kosoy et al.\textsuperscript{11} worked on providing continental ancestry and characterized a set of 128 AIMs. The markers were chosen for informativeness, genome-wide distribution, and genotype reproducibility from 825 individuals. There are several ancestry estimation software that quantify genetic variation of admixture between populations using high-throughput sequencing data, such as the models used in the programs STRUCTURE\textsuperscript{11}, FRAPPE\textsuperscript{14}, TESS\textsuperscript{15}, and Admixture\textsuperscript{16}.

However, many existing studies on race/ethnicity identification are restricted by sample size or biased by sample selection. For example, the evaluations for Eriksson et al.\textsuperscript{17} were done only on the European population in America. Yaeger et al.\textsuperscript{10} specifically focused on African Americans born in the United States and in Africa. A total of 825 individuals were examined by Kosoy et al.\textsuperscript{11} covering a wider range of individuals, but the sample size is still limited. Aggregating data from multiple sources could significantly improve the power of the study in race/ethnicity identification. However, direct sharing of labeled patients’ genetic information for data mining would violate the policies concerning patient privacy\textsuperscript{18}. Besides the privacy concerns in data mining phase, the same issues are also associated with the testing phase, where a researcher needs to identify the ethnicity through individual’s genomic data, but without compromising the patient’s privacy.

Regarding the privacy concern, human genomic data must be handled carefully to avoid disclosure of sensitive patient information to unauthorized parties. Previous studies\textsuperscript{19–23} have demonstrated several privacy risks regarding to human genomic data. For example, Homer et al.\textsuperscript{24} discovered that the presence of an individual in a case group can be reliably determined (known as a re-identification attack) from the allele frequencies using an individual’s DNA profile, which can be acquired, for example, from a single hair or a drop of blood. The biomedical community has recognized the importance of privacy and data protection for genomic projects\textsuperscript{25}. Many privacy and security technologies, e.g., differential privacy (DP)\textsuperscript{26}, homomorphic encryption (HME)\textsuperscript{27–30} and secure multiparty computation (SMC)\textsuperscript{31,32} have advanced in protecting biomedical data\textsuperscript{33–39}. Among them, DP solutions will alter data to make it difficult to identify information to a particular individual, and DP might also render outputs useless\textsuperscript{25}. HME and SMC (e.g. based on garbled circuits and secret sharing) hold the promise of secure general-purpose computing in the cloud but existing solutions are too computationally cumbersome to be used for complex big-data analysis. In addition, efficient SMC solutions exist (e.g. based on secret sharing and arithmetic circuit) but are domain-specific, thus inappropriate for exploratory analyses that need constant tuning. Distinct from some of the existed tools, our pipeline will not only allow users to perform ethnicity detection of a patient, but also provide secure protection of the subject’s information (Figure 1). We chose to use Intel\textsuperscript{20} Software Guard Extensions (Intel\textsuperscript{20} SGX)\textsuperscript{40}, which is a set of new CPU instructions that can be used by applications to set aside private regions of code and data. Intel\textsuperscript{20} SGX allows application to protect sensitive data from unauthorized access or modification and enables application to preserve the confidentiality without disrupting the software system. Both the labeled data and the unlabeled data from different sources will be protected during the process (Figure 1). Clients can only obtain the encrypted results (e.g., estimation of individual admixture) but cannot access the training data deposited by other institutions.

**Materials and Method**

In this paper, we will focus on the secure estimation of individual admixture using genomic data. We will first introduce the methods used for selecting AIMs and identifying individual admixture followed by the details about Intel\textsuperscript{20} SGX frameworks.

*Selection of AIMs:* The selection of the most discriminative AIMs can significantly improve the efficiency of prediction for identifying individual admixture\textsuperscript{10}. In this paper, we compute and sort the mutual information\textsuperscript{12} between multiple population groups to select the most discriminative bi-allelic SNPs (used as AIMs in this study).
More specifically, the markers that are informative in one collection of source populations are generally informative in others\textsuperscript{10}. Therefore, we apply statistical methods that use multilocus genotypes and population allele frequencies to represent the average frequency of allele at a certain locus, and assign informativeness to no-admixture model based on the conditional entropy of a random population given the knowledge of the genotype. Next, the informativeness is sorted from the largest to the smallest.

**Identifying individual admixture:** Our pipeline determines an individual’s ethnicity based on the method developed by Tang et al.\textsuperscript{14}. The Estimation Maximization (EM) algorithm proposed by this model demonstrates increased robustness and comparable efficiency when compared to existing maximum likelihood (ML) model\textsuperscript{12} and Bayesian MCMC method\textsuperscript{13}. This estimation allows for uncertainty in ancestral allele frequencies, provides an advantage toward separating an admixture population at an individual level, which they referred to as “individual admixture” (IA), and could achieve extensive stimulations to produce point estimates\textsuperscript{14}. The goal is to estimate IA for the admixed individuals, $Q_i = (q_{i1}, ..., q_{ik})$ and $i = 1, ... I_0$, where $Q_i$ is the individual admixture fraction of the $i$-th individual; $I_0$ is the total number of individuals to be identified; $K$ is the number of ancestries. The allele frequencies of different ancestral markers are denoted as $P$. A likelihood function given an unobservable variable $Z_{ima} \in \{1, ..., K\}$ can be expressed as

$$l(G, Z\mid P, Q) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{k=1}^{K} 1(G_{ima} = l, Z_{ima} = k) \log(p_{mik} q_{ik})$$

where $1(\cdot)$ is an indicator function; $p_{mik}$ is the frequency of allele $l$ at marker $m$ for the $k$-th ancestry; $G_{ima}$ is the allele for the $i$-th individual at the $a$-th allele of marker $m$. An EM algorithm can be implemented to efficiently estimate the parameters $P$ and $Q$. Convergence is declared when the difference in the estimates of $Q$ and $P$ fall below a small threshold\textsuperscript{14}.

**Secure computation methods:** To mitigate the privacy risk while supporting scientific discovery, security researchers have developed many theoretical frameworks. However, even the best-known methods based on HElib\textsuperscript{41} for homomorphic encryption or FlexSC\textsuperscript{42} for garbled circuit-based secure multiparty computation are not practical enough to handle large-scale genomic data analysis. These real challenges motivate the development of new solutions. Intel recently announced a new Software Guard eXtensions (SGX)\textsuperscript{40} architecture for their next generation CPUs, which shed light to novel solutions to above mentioned challenges using a hybrid software-hardware framework. Figure 2 shows some of the major conceptual functions of the SGX architecture (need to be developed on a case-by-case manner depending on the application). As a built-in feature in the Intel\textsuperscript{10} Skylake family Central Processing Units (CPUs), SGX-enabled devices can be found in most recently released computing platforms (e.g., laptops, desktops and servers). Intel\textsuperscript{10} SGX framework provides a cost-effective solution to achieve affordable secure computation in terms of both complexity and finance concerns. For example, a pioneer study of SGX-based MapReduce framework for distributed high-performance computations\textsuperscript{43} shows a negligible overhead of 8% to achieve read/write integrity using SGX. This is a significant advantage of using SGX in comparison to other secure computation scheme like homomorphic encryption or garbled circuits, which usually increase the complexity thousands of times over. Furthermore, an SGX-enabled machine only cost as low as a few hundred dollars. In SGX, a protected area in CPU, which is usually referred to as enclave, is dedicated to execute sensitive codes and compute sensitive data in a secure manner, where any interfere from software outside the enclave are prohibited by the SGX hardware. Therefore, both data confidentiality and integrity can be achieved with a proper systemic design of SGX applications. SGX is resilient to both software level attacks (e.g., malicious operating system, etc.) and hardware level attacks (e.g., for memory, hard disk, network etc.). Some preliminary studies\textsuperscript{44-45} have revealed the possibility of SGX to significantly enhance the security and privacy of many applications. However, most of them are based on simulation study and none of them has tackled genomic data security and privacy in a real SGX-enabled computing platform.
To ensure the security of the whole system, an SGX framework requires the adoption of industry-standard cryptographic primitives and systemic implementation of several key steps, as shown in Tables 1 and 2, respectively. Taking advantage of this novel architecture, we developed a PREMIX framework for privacy-preserving estimation of individual admixture in this paper.

Table 1. Summary of cryptographic primitives to be adopted in the design of SGX application.

<table>
<thead>
<tr>
<th>Cryptographic primitives</th>
<th>Description</th>
<th>Security</th>
<th>Industry standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Encryption Standard (AES) in Galois Counter Mode (GCM)</td>
<td>Authenticated encryption, which provides simultaneous protection of data confidentiality and authenticity.</td>
<td>128 bits</td>
<td>NIST SP 800-38D guideline</td>
</tr>
<tr>
<td>Elliptic Curve Diffie–Hellman (ECDH)</td>
<td>A key agreement protocol based on Elliptic Curve Cryptography (ECC) to establish securely shared symmetric key for AES over an insecure channel.</td>
<td>256 bits</td>
<td>NIST SP 800-56A guideline</td>
</tr>
<tr>
<td>Elliptic Curve Digital Signature Algorithm (ECDSA)</td>
<td>An ECC based digital signature scheme to ensure the source of data is as claimed.</td>
<td>256 bits</td>
<td>FIPS Pub. 186-3 guideline</td>
</tr>
</tbody>
</table>

Table 2. Key steps and their corresponding cryptographic primitives for SGX application to achieve efficient and trustworthy computation.

<table>
<thead>
<tr>
<th>Key steps</th>
<th>Description</th>
<th>Cryptographic primitives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote attestation</td>
<td>Securely provision an enclave from an authorized user who is outside the computing platform</td>
<td>AES-GCM ECDH ECDSA</td>
</tr>
<tr>
<td>Data provision</td>
<td>Securely transfer sensitive data into enclave</td>
<td>AES-GCM ECDH ECDSA</td>
</tr>
</tbody>
</table>

Results

Study Subjects and Ancestral Population: We have used the phase 3 release (May 2013) of the 1000 Genomes project data to identify SNPs. Datasets from 1000 Genome project include genetic variation across diverse populations from Europe, Asia, Africa and the Americas. The present 1000 Genome data contain 2504 samples from 26 populations which can be categorized into five super-populations: East Asian (EAS), South Asian (SAS), African (AFR), European (EUR), and American (AMR). The global allele frequencies for each super-populations were calculated by counting the AC (“Total number of alternate alleles in called genotypes”) and AN (“Total number of alleles in called genotypes”) for all the individuals from a particular super population and using that to calculate the allele frequencies.

Experimental setup: We use both simulation and real data to test the performance of our algorithm. For simulation experiment, we follow the setup used by Tang et al. Assume there are two ancestral populations X and Y. The simulated data consist of 500 admixed individuals, and 250 individuals from each of two populations as the training data. Since there are two groups, the IA vector is a scalar, and we sample it from a mixture model of Gaussian and uniform distributions. We select the SNPs that the differences of its allele frequencies from two parties are bigger than a δ-value 0.3. Conditioning on the SNPs and the IA vector, we can randomly generate the 1000 individuals for simulation. From 1000 genome data, we used two populations: ACB (African Caribbeans in Barbados) and TSI (Toscani in Italia), and each of them contains 96 and 107 subjects. We extracted the first 31,000 SNPs of their 22nd chromosomes. All of these SNPs will be processed by the informativeness algorithm to choose the top SNPs for computing the IA vector of the admixed individuals.

We implemented the proposed PREMIX server and client on two machines: the server is an Intel® Xeon core E3-1275 v5 with Intel® SGX support and 64 GB memory; the client machine is Intel® Core i7-6820HQ CPU and 48GB Memory.

The experiments are designed to focus on the following three aspects: (a) computational complexity comparison between secure SGX-based C++ implementation vs insecure C++ implementation; (b) the simulation data results; (c) the real data results using different number of top informative AIMs.
**Experimental results:** Table 3 shows the key steps and total running times of PREMIX using secure SGX implementations with encrypted remote data and insecure C++ with local data. We tested four different data sizes, and all of the results in Table 3 are the averaging values over 10 trials. From the results, we can see that, there is no significant different between the two frameworks in implementing EM algorithm, but the total running times of secure SGX is a little slower than those of insecure C++. The additional overhead in the total running times of secure SGX is due to the data encryption, attestation, data transfer and analysis over encrypted data.

Table 3. Comparison of the running times of computing PREMIX between secure SGX-based C++ implementation vs. insecure C++ implementation. Here $I_0$ and $I$ are the number of admixed individuals and the total individuals, respectively. The unit of all running times is second.

<table>
<thead>
<tr>
<th>$I_0/I$</th>
<th>Secure SGX-based C++</th>
<th>Insecure C++</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Client data encryption</td>
<td>EM algorithm</td>
</tr>
<tr>
<td>250/500</td>
<td>&lt;0.001</td>
<td>1.827</td>
</tr>
<tr>
<td>500/1500</td>
<td>0.006</td>
<td>3.613</td>
</tr>
<tr>
<td>750/1500</td>
<td>0.008</td>
<td>6.937</td>
</tr>
</tbody>
</table>

Simulated data is used for evaluating the PREMIX framework. We suppose that the labeled data with ethnicity information are non-admixed and there are two populations. Each individual in the labeled data set is sampled from one of the two populations. The number of SNPs used in our simulation is 200. The iteration rounds of the EM algorithm are set to 200.

Table 4. Performance based on simulated data.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>RMSE</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.117</td>
<td>0.015</td>
</tr>
<tr>
<td>0.4</td>
<td>0.089</td>
<td>0.006</td>
</tr>
<tr>
<td>0.5</td>
<td>0.080</td>
<td>0.010</td>
</tr>
<tr>
<td>0.6</td>
<td>0.070</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 5. Percentage of correctly identified individuals using different number of AIMs.

<table>
<thead>
<tr>
<th>$M$</th>
<th>Percentage of correct identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>81%</td>
</tr>
<tr>
<td>20</td>
<td>93%</td>
</tr>
<tr>
<td>50</td>
<td>95%</td>
</tr>
<tr>
<td>100</td>
<td>95%</td>
</tr>
<tr>
<td>200</td>
<td>95%</td>
</tr>
<tr>
<td>500</td>
<td>97%</td>
</tr>
</tbody>
</table>

Table 6. Percentage of correctly identified individuals using different labeled data sizes.

<table>
<thead>
<tr>
<th>Labeled data size</th>
<th>Percentage of correct identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>93%</td>
</tr>
<tr>
<td>40</td>
<td>95%</td>
</tr>
<tr>
<td>60</td>
<td>95%</td>
</tr>
<tr>
<td>80</td>
<td>95%</td>
</tr>
<tr>
<td>100</td>
<td>95%</td>
</tr>
</tbody>
</table>

Table 4 is the results of PREMIX using simulated data. Since we know the ground truth under simulation environment. We can compute the root mean square error (RMSE) and the bias. Table 4 shows that the RMSE will less than 0.1 with no more than 0.01 bias, if the $\delta$ value of two groups is bigger than 0.3.

![Figure 3](image1.png)  
(a) Two populations  
![Figure 3](image2.png)  
(b) Three populations

**Figure 3.** Estimated individual admixture using the proposed PREMIX framework.

For the real data, we split each population into two groups, where the first group is used as admixed individuals to be predicted and the second group is used as the labeled data. Specifically, the individuals from ACB and TSI
populations were included in our experiment. For both populations, the first 50 individuals are viewed as admixed individual, and the rest individuals from both populations (i.e., 46 ACB individuals and 57 TSI individuals) are used as labeled data. In addition, we suppose these labeled data were from two sources to simulate a secure collaboration scenario. The SNPs were screened based on their informativeness, where the top M SNPs will be selected for the next step in PREMIX framework. For the real data, since there is no ground true of IA vector, the admixed individual will be classified to the population based on their maximum estimated IA component. Tables 5 and 6 depict the percentage of correctly identified individuals using different number of AIMs and different labeled data sizes, respectively. In Table 5, we can see that the percentage of correct identification increases as the number of AIMs M increases. Based on our experiments, a high percentage of correct identification can be achieved with 50 or more AIMs. Moreover, Table 6 shows the framework can achieve a relative accurate identification performance with as few as 40 labeled data.

Figure 3 (a) shows the estimated IA vectors for two populations. The first 50 individuals are from the ACB population, and the second 50 individuals are from TSI population. We can see that the PREMIX can successfully identify the ethnicities of most individuals in both groups.

To further evaluate the performance of the PREXIM, we included a third population CHS (Southern Han Chinese) in our experiment. In Figure 3 (b), the first 50 individuals are from the ACB; the second 50 individuals are from CHS; and the final 50 individuals are from TSI. We see that there is some performance degradation of the proposed PREXIM framework in identifying more than two populations.

**Discussion and Limitation**

The main contribution of this paper is to introduce a new hybrid solution (i.e., Intel® SGX) using both hardware and software to enable efficient and privacy-preserving estimation of individual admixture. The proposed PREMIX framework can protect the privacy of sensitive genomic data with ancestry information, as well as the privacy of the data users, who would like to identify their individual admixture. Due to the adoption of strong security protection primitives, multiple data owners can collaborate on the study to improve the estimation performance without sacrificing individual data privacy. In the proposed framework, we provided both a secure feature selection module based on informativeness of SNPs and a secure EM based maximum likelihood estimator to achieve both computational efficiency and estimation accuracy. Our experimental results demonstrated the advantage of secure collaboration in identifying individual admixture.

There are several limitations in this study. First, even it can well protect the data privacy, the SGX hardware is vulnerable to Denial-of-service (DoS) attack; however, the data privacy will not be compromised under this attack. Second, proposed method was only evaluated through limited data sets (i.e., simulated data and 1000 genome data) in this pilot study. The use of Human Genome Diversity project (HGDP) could improve the impact of this study and provide better performance assessment. In addition, the proposed method relies on an EM-based maximum likelihood estimator, which can only support a small number of populations. Recently, many advanced ethnicity identification programs have been developed particularly for genome-wide SNP data. For example, TESS, an updated version of the spatial ancestry estimation program TESS, which combines matrix factorization and spatial statistical methods. TESS estimates ancestry coefficients with comparable accuracy and fast run-times, and can be used to perform genome scans for selection, separate adaptive from non-adaptive genetic variation using ancestral allele frequency differentiation tests. AncestryMapper assigns each individual analyzed with a genetic identifier, referred as Ancestry Mapper Id (AMid) which corresponds to its relationship to the HGDP reference population. TreeMix follows a tree-based approach with branches built by maximum likelihood lengths and migration weights. Population was identified by searching through the space of possible graphs with optimized the branch lengths and weights. FastSTRUCTURE is a recent modification of the popular model STRUCTURE which provides a faster approximate inference using a variational Bayesian framework and poses the problem of computing relevant posterior distributions as an optimization problem. The software identifies the number of populations represented in a dataset with heuristic and new hierarchical prior to detect weak population structure in the data. Among a large population, algorithm such as Eagle detects association analysis of rare variants from a large population cohorts based on genotyping arrays using long-range phasing (LRP) to rapidly phase segments of genome identical-by-descent (IBD) with closely or distantly related individuals. Eagle runs two iterations of fast approximate Viterbi decoding using a simple diploid analog of the Li-Stephens HMM to allows phasing of segments lacking IBD to ensure accurate results. The development of trustworthy computation framework to support advanced methods in race/ethnicity identification warrants the further investigation along this line. Finally, the limited secure memory (~
96 MB) in SGX restricts the algorithm to process a huge amount of data concurrently. In the next step, we will optimize the secure memory usage to improve the data processing capacity of the proposed method.

Acknowledgment

F.C., M.D., Y.L. and S.W. contributed to the majority of the writing. F.C. and S.W. designed the method and devised the simulation experiments. F.C., Y.L., S.D. and S.W. implemented algorithms. X.J. provided detailed edits. S.W., X.J. and H.T. provide critical suggestions to improve the paper. This work has been supported by the NHGRI (R00HG008175, R01HG007078), NIGMS (R01GM114612), NLM (R00LM011392, R21LM012060), NHLBI (U54HL108460), NIH bioCADDIE (1U24AI117966-01).

References


40. Intel® Software Guard Extensions (Intel® SGX) [Internet]. Available from: https://software.intel.com/en-us/ssa-extensions/intel-sgx


42. Wang X. A Flexible Efficient Secure Computation Backend computer program (College Park, Department of Computer Science University of Maryland) [Internet]. 2015 [cited 2016 Feb 1]. Available from:
https://github.com/yhuang912/FlexSC


Abstract

Chronic disease affects patient quality of life through symptoms of the disease and the work of receiving treatment. While the effects of illness are well investigated, the burden of treatment is not commonly studied or monitored. We developed a method to quantify one dimension of the burden of treatment based on patient encounters with the healthcare system. We applied this method to a population of stage I-III breast cancer patients. As hypothesized and observed, stage III patients had more appointments, spent more time in clinic, and spent more time admitted to the hospital in the first 18 months after diagnosis compared to stage I and II patients. Future work will evaluate the reproducibility and generalizability of this method for quantifying burden of treatment across other clinical settings and chronic diseases. This approach could enable identification of high-risk groups that could benefit from interventions to decrease patient work and improve outcomes.

Introduction

Chronic illness is detrimental to a patient’s quality of life both because of the symptoms of the illness itself as well as the burden of treatment needed to combat the illness. The complexity of medical care today makes it difficult for healthcare providers to monitor a patient’s capacity to receive care even though treatment overburden can impact disease outcomes. A high burden of treatment can cause lower compliance in patients with chronic diseases, thus increasing other complications. Patients with high burden of treatment tend to have lower satisfaction scores. Excessive treatment can also lead to wasted resources for the medical center, and contribute to a patient’s financial toxicity. Physicians who practice minimally invasive care assess burden and tailor treatment plans that give a patient the maximum likelihood of recovery while taking into consideration the patient’s limitations. To improve the effectiveness of this paradigm, providers and healthcare systems need reliable ways to identify overburdened patients and patient populations.

Cancer patients often undergo intense, multi-modal treatments resulting in diminished quality of life. Figure 1 shows a typical two-week schedule for a breast cancer patient undergoing adjuvant chemotherapy following surgery. This schedule includes 11 appointments over 7 unique days and a total clinic time of 9.5 hours. Additional appointments needed to address complications and comorbid conditions compound the treatment burden for cancer patients. A nationwide study showed that among patients undergoing chemotherapy and radiotherapy, 28% had to schedule appointments to treat side effects, 77% had to arrange for caregivers to accompany them to their appointments, and 43% had some impact to their professional lives.

Figure 1. Example of a breast cancer patient’s schedule over a two-week period. This patient had 11 appointments over 7 unique days and a total clinic time of 9.5 hours between July 11 and August 1. MO = Medical Oncologist, NP = Nurse Practitioner.
Burden of treatment measures focus on the impact of the acts of receiving care. Previous research on the topic of treatment burden have focused on qualitative methods to describe factors contributing to patient work\textsuperscript{13}. By surveying patients, researchers identified a taxonomy of factors that contribute to the burden of treatment including healthcare tasks and situational factors that exacerbate a patient’s work\textsuperscript{14,15}. Tran et. al. also developed and validated the Treatment Burden Questionnaire (TBQ), a survey instrument designed to measure patient burden\textsuperscript{5,14}. Disease specific questionnaires have also been developed to assess the burden of treatment in specific chronic conditions such as chronic heart failure\textsuperscript{16} and end stage renal disease requiring dialysis\textsuperscript{17}. While these qualitative methods are effective in defining treatment burden, they do not enable population studies of burden or automated monitoring to identify overworked patients who may need intervention.

Figure 2. Relative completeness and structure of data elements related to factors of treatment burden that may be amenable to automated extraction and quantification from the electronic health record.

Research using electronic health record data is most effective when that data is structured and complete\textsuperscript{19}. Figure 2 displays several dimensions of treatment burden identified by Tran et. al.\textsuperscript{13} and Eton et. al\textsuperscript{14} evaluated by approximate completeness and structure of data in the electronic health record. Treatment burden elements in the top-right quadrant of Figure 2 such as financial costs and clinic visits are both highly structured and highly complete by virtue of their relation to billing. These top-right elements are more reliable measures than elements in the bottom-left that are unstructured and less available. For example, a patient’s exercise program is less likely to be captured in an EHR but may be recorded in other systems managed by the patient outside the EHR. While healthcare institutions may differ in how these factors of treatment burden are recorded, many of these data elements are available across different implementations of clinical information systems.

This study proposes a method to quantify one of the more reliable contributors to treatment burden: clinical encounters. While appointments only contribute to part of a patient’s overall treatment burden, time and effort spent coordinating, traveling for, and waiting for care were among the most commonly mentioned contributors to burden in patient surveys\textsuperscript{13}, and the factors most highly correlated with the overall TBQ score\textsuperscript{5}. Hospital admissions are also of particular interest to patients and hospital administration. Cancer related readmissions are often unexpected and the most frequent among patients with private insurance\textsuperscript{15}. Although inpatient time was typically not mentioned in previous burden of treatment literature, admissions are disruptive to patients and their caregivers. To our knowledge, no previous studies have used data from electronic health records to assess treatment burden in patient populations. By using appointment and admission data, our evaluation of treatment burden is reliable, reproducible, and scalable given accurate electronic records.

To evaluate our method, we chose to investigate patients with stage I-III breast cancer. We hypothesized that patients with stage III cancer would have a higher treatment burden than patients with stage I or stage II cancer given the more aggressive therapy they receive for higher risk disease. Additionally, we anticipated that burden of
treatment would be greatest in the first few months after the date of cancer diagnosis, and that this burden would decrease over time.

**Methods**

To study the use of these treatment burden metrics, we applied them to a population of breast cancer patients at Vanderbilt University Medical Center (VUMC). The goal of this analysis was to investigate whether there was differentiation in number of appointments, total time spent in clinic including wait time, and total inpatient length of stay between breast cancer patients with stage I, stage II and stage III disease. We also identified outliers who had abnormally high treatment burden for their stage.

**Study population**

The patient cohort for this study was chosen from the VUMC Cancer Registry since those were the patients who were diagnosed and receive all or part of their first course of treatment at our institution. We collected 18 months of encounter information from all stage I-III breast cancer patients who were diagnosed over a 17-year period between January 1, 1998 and June 1, 2014. In order to facilitate comparison between sub-populations of patients that received the majority of their care at our institution, we included only patients with at least three appointments from both a Vanderbilt medical oncologist and a Vanderbilt surgical oncologist. We determined which appointments were with a medical or surgical oncologist by mapping their national provider identification number with their specialty in the national patient identifier (NPI) data dissemination file.

Among the 8161 patients with breast cancer in the VUMC Cancer Registry, 5661 had a date of diagnosis between January 1, 1998 and June 1, 2014. Among these, 4152 patients had stage I-III disease at diagnosis. After collecting 18 months of appointments after the date of diagnosis for these patients, we found that 904 had more than three appointments with a medical oncologist and a surgical oncologist at VUMC (Figure 3).

**Data Considerations**

To get a holistic view of each patient’s treatment burden, we included all patient appointments, including appointments in non-cancer related departments, and all hospital admissions. We extracted outpatient appointment data from the Epic scheduling system that had been in use at VUMC since 1997. Hospital admission and discharge data was extracted from the Medipac admission, discharge, transfer (ADT) system with data going back to 1984. To approximate the total time spent in clinic including waiting time, if a patient had more than one appointment in a day, we calculated the time from the beginning of the first appointment to the end of the last appointment of the day. We used the schedule appointment duration as an estimate of the time the patient spent in clinic understanding that true appointment durations could vary.

**Analysis**

We compared distributions of the total time spent in clinic, the number of appointments, and the number of admissions over 18 months by stage. We used an ANOVA test to see if there was a significant difference between stages I-III for each of the means of these three metrics. We also compared the estimated time spent in clinics by month over 18 months by stage. The Vanderbilt Institutional Review Board approved this study and granted a waiver of consent since we analyzed a large population of patients in aggregate.
Results

Table 1 summarizes the clinical encounter burden by stage. Among the 904 patients in the final cohort, 419 had stage I, 337 had stage II, and 148 had stage III disease. Across all stages, the average patient in our cohort had 67 appointments on 44 unique appointment days, spent 3.6 hours in clinic per month, and was admitted for 1.9 days over the 18 months after diagnosis. On average, Stage III patients had the greatest number of appointments (92), the greatest amount of time spent in clinic per month (5.4 hours), and the greatest number of hospitalized days (3). Stage II patients had the second greatest totals in each of these parameters and stage I patients had the least. The ANOVA tests for all of the metrics were significant with p-values less than .001.

Table 1. Summary of clinical encounter burden for breast cancer patients by stage with ANOVA p-values comparing the difference between stage I, stage II, and stage III.

<table>
<thead>
<tr>
<th>Breast Cancer Stage</th>
<th>Stage I</th>
<th>Stage II</th>
<th>Stage III</th>
<th>Stage I-III</th>
<th>ANOVA p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>419</td>
<td>337</td>
<td>148</td>
<td>904</td>
<td>N/A</td>
</tr>
<tr>
<td>Number of appointments</td>
<td>53 (8-164)</td>
<td>74 (9-254)</td>
<td>92 (15-217)</td>
<td>67 (8-254)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Unique appointment days</td>
<td>37 (7-126)</td>
<td>47 (4-145)</td>
<td>57 (6-124)</td>
<td>44 (4-145)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Hours of appointment time</td>
<td>26 (3-128)</td>
<td>48 (3-195)</td>
<td>62 (8-152)</td>
<td>40 (3-195)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Hours spent waiting between appointments</td>
<td>17 (0-97)</td>
<td>28 (5-95)</td>
<td>36 (3-92)</td>
<td>24 (0-97)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Hours spent in clinic</td>
<td>43 (3-163)</td>
<td>76 (6-247)</td>
<td>98 (12-233)</td>
<td>64 (3-247)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Hours spent in clinic per month</td>
<td>2.4 (.17-9)</td>
<td>4.2 (.33-14)</td>
<td>5.4 (.67-13)</td>
<td>3.6 (.17-14)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Number of unique admissions</td>
<td>.51 (0-10)</td>
<td>.87 (0-8)</td>
<td>1.1 (0-7)</td>
<td>.74 (0-10)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Total inpatient length of stay (days)</td>
<td>1.3 (0-41)</td>
<td>2.3 (0-33)</td>
<td>3.0 (0-29)</td>
<td>1.9 (0-41)</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Figures 4A and 4B. Distribution of total number of appointments over 18 months by breast cancer stage (ANOVA p-value < .001), and total length of stay over 18 months by breast cancer stage (ANOVA p-value < .001). The dark line for each boxplot represents the median and the colored box represents the interquartile range (IQR) (25th to 75th
percentile). The “whiskers” extend to 1.5 times the IQR or to the minimum or maximum value, whichever is closer. Any data points outside the whiskers are outliers and are represented individually as circles.

The boxplot in Figure 4A further shows that the median and interquartile ranges of the number of appointments is greatest for patients with stage III disease followed by stage II and stage I. There were 5 outliers for stage I patients with an unusually high number of appointments while there were 2 outliers for stage II patients. Figure 4B shows the distribution of total length of hospitalized days in the first 18 months after diagnosis. The median patient for stage I and stage II was admitted for one day or less while the median stage III patient was admitted for about two days. Among stage I patients, 67% had no admissions, while only 49% of stage II and 41% of stage III patients had no admissions. With so many patients clustered near zero, there were many outliers for all stages.

**Figures 5A and 5B:** Distribution of total time spent in clinic over 18 months by breast cancer stage (ANOVA p-value <.001), and average total time spent in clinic per month over 18 months by breast cancer stage.

Figure 5A shows that the median time spent in clinic was greatest for stage III patients followed by stage II and stage I. There are 26 outliers for stage I patients in time spent in clinic and 2 for stage II patients. Over the course of 18 months after diagnosis, the average total time spent in clinic per patient decreased (Figure 5B). In the first month of treatment, stage I patients spent on average 8 hours in clinic while stage II and III patients spent 10 and 12 hours respectively. All three stages saw a decrease in time spent in clinic in the second month but then had an increase in the third and fourth months. In each month after diagnosis, the average time spent in clinic was greater for stage III than stage II or stage I breast cancer patients, although this difference was not statistically significant for any given month.

**Discussion**

We have developed a simple method for quantifying an important aspect of treatment burden using clinical encounter data derived from the outpatient scheduling and inpatient ADT systems. The results of the ANOVA tests from Table 1 confirm our hypothesis that there is significant separation between stages I-III breast cancer patients with respect to the number of distinct appointments, total time spent in clinic, and total number of hospitalized days. We also observed that treatment burden diminished over time in the 18-month period following diagnosis (Figure 5B). This decrease is expected since the most intense portion of treatment for early stage breast cancer is usually completed in the first 12-14 months after diagnosis. Our metric is concordant with this pattern of treatment.
This approach for assessing burden of treatment is simple and generalizable to other healthcare organizations. Since Stage 1 Meaningful Use was enacted, healthcare systems are incentivized to maintain accurate patient encounter records\textsuperscript{21}. Therefore, any healthcare organization could use scheduling data to approximate patient burden. On the other hand, some limitations impede direct comparisons between our study population and those at other institutions. Other institutions may record appointments differently than VUMC. For example, another healthcare system with the patient in Figure 1 could have just one appointment for lab, medical oncologist, and infusion. Meanwhile at VUMC, each of those encounters is recorded as a separate appointment. Comparing patient populations within that institution would still be possible but comparing populations across institutions might be challenging.

Our method of determining the amount of time spent in clinic is a gross estimation. It does not take into consideration situations where patients arrive early, appointments start late, patients leave between appointments, or appointments end earlier than the time allotted. Using data from systems such as VUMC’s outpatient whiteboard would enable detailed analysis of patient whereabouts\textsuperscript{32}. For even more granular data, healthcare systems have used real time locator systems (RTLS) to pinpoint the location of patients as they move through the medical center\textsuperscript{23}. These advanced techniques would provide precise data about patient burden during patient encounters but would be difficult to generalize to other institutions that do not have infrastructure for patient tracking.

Future work in developing quantitative measures for assessing treatment burden includes incorporating additional factors that influence patient outcomes identified in previous literature. To more accurately capture the patient experience related to appointments, we plan to incorporate commute time into their burden assessment by adding the time to drive from their home address to the clinic address before and after appointments. We will use other structured data such as medication prescriptions to determine the frequency of home medication use, and billing information to approximate other medical encounters not captured directly as appointments. There is also potential for natural language processing of notes to capture other provider recommended activities crucial to outcomes such as exercise or diet changes. Complementary to the need for an accurate assessment of a patient’s treatment burden is the determination of a patient’s capacity for treatment. With burden and capacity, we can compare outcomes for patients for whom burden exceeds capacity against those who receive care within their means.

There are multiple applications where a data driven and quantitative measurement of treatment burden could impact patient, provider, and system level decision-making. Using the electronic health record to characterize burden is much like phenotype classification, where researchers use electronic health data to group similar patients for more personalized care delivery\textsuperscript{24}. Understanding treatment burden can aid in delivering the right amount of care that is prioritized to what each individual can handle. When patients are newly diagnosed with stage I-III breast cancer, Figure 5B could help them anticipate how much time they will need to devote to coming to receive care and how much time they will need to take off from work. Likewise, Figure 4B could help them understand how much time patients like them are admitted in the hospital. Future work could further divide our cohort into patients who chose different treatment paths such as prophylactic contralateral mastectomy followed by reconstruction compared to lumpectomy with radiation therapy. Showing the treatment burden of similar patients could help educate patients about treatment options and their trade-offs.

Our analysis could also help with provider monitoring of patient burden. Our work looking at treatment burden is analogous to using electronic health records to predict risk for readmission. In readmission prediction, clinicians use social and clinical factors to identify patients who are high risk for readmission, which is costly to the medical center and harmful to the patient\textsuperscript{25}. Similarly, treatment burden factors can identify patients at risk for non-compliance and wasted resources which are also costly to the medical center and harmful to the patient. The boxplot in Figure 4A is a good example of how healthcare systems can use appointment data as a proxy for treatment burden to identify outliers. In our population, a provider would notice that there are five stage I patients who had around 150 appointments in an 18-month period. A patient care team for one of those patients could investigate whether the appointments are appropriate. If that care is necessary, the healthcare system may look into ways to help ameliorate burden such as home visits or transportation assistance. Additionally, these types of data demonstrate the potential risks in pursuing alternative payment models such as bundled payment models. Administrators and providers could use these tools to help monitor for patients at high risk for high care utilization.

Focusing on outliers in appointment burden could also identify opportunities for improved care coordination and more convenience for the patient. For stage I patients in Figure 5A, there are many more outliers for total time spent in clinic than for count of appointments in Figure 4A. Since Figures 4A and 5A visualize the same population of patients, an increase in the number of outliers means that some patients who are outliers in time spent in clinic either have longer appointments or more time between their appointments. These patients would be candidates for care.
coordination interventions such as arranging their appointments closer together on the same day, or assigning them to a medical home clinic.

There were several potentially confounding factors that we attempted to control for in our cohort selection. The first and most significant is that we had to determine which patients received their first course of treatment at VUMC. Although the cancer registry had information about which providers saw patients in the registry and what institutions they were from, the availability of that data was inconsistent. We decided on a data driven approach where we defined patients as having received their first course of treatment at VUMC if they had at least three appointments with both a medical oncologist and surgeon from VUMC in the first 18 months of treatment. This constraint cut our cohort by more than 75%, but enacting the constraint was necessary to ensure that analysis focused on patients receiving care where we had more complete data on their encounters. We limited our analysis to stage I-III patients for a similar reason. Stage 0 and incurable stage IV patients have very different patterns of care making them less comparable in the first 18 months of treatment to stage I-III patients. We chose 18 months as the interval for analysis since a typical course of treatment occurs within that time frame and there is a very low risk of disease recurrence during this time.

Another limitation of this study is that we were not able to address missing data from patients lost to follow-up during the 18-month time span, or for care patients received outside of our institution. Furthermore, during the 17-year period where we observed our patients, there were changes in the way appointments were recorded in the system. There was a gradual increase in appointments per patient due to an increase in departments using the scheduling system. The effect of this increase in appointment capture should have been minimal on the analysis since they would be equally distributed across patients in the different stages.

Conclusion

With increasing interest in care coordination and value based care, there are already many reasons to limit unnecessary treatment for chronic diseases. Evidence is mounting that appropriate care for patients with a given disease may differ depending on their capacity to handle the burden of treatment. Our study of patient encounters is the first step toward a comprehensive and automated method to assess treatment burden applicable to disease population comparisons. In comparing patients with breast cancer, we found significant differentiation in appointments and admissions for patients with stage I, II, and III disease, as well as a downward trend in amount of time spent in outpatient appointments over time after diagnosis. Future work includes incorporating more factors that influence treatment burden and comparing treatment burden across different types of chronic disease. By better understanding burden of treatment, we can begin to deliver precision medicine not only based on genetic makeup and disease phenotypes, but also on the patient’s capacity to comply with treatment plans in order to maximize the likelihood for improved outcomes.

References

8. Shippee ND, Shah ND, May CR, Mair FS, Montori VM. Cumulative complexity: a functional, patient-


An Empirical Study for Impacts of Measurement Errors on EHR based Association Studies

Rui Duan, M.S.¹, Ming Cao, M.S.², Yonghui Wu, Ph.D.³, Jing Huang, Ph.D.²,
Joshua C Denny⁴,⁵, Hua Xu, Ph.D.³, Yong Chen, Ph.D.¹

¹Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA
²School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA
³School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
⁴Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
⁵Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

Abstract

Over the last decade, Electronic Health Records (EHR) systems have been increasingly implemented at US hospitals. Despite their great potential, the complex and uneven nature of clinical documentation and data quality brings additional challenges for analyzing EHR data. A critical challenge is the information bias due to the measurement errors in outcome and covariates. We conducted empirical studies to quantify the impacts of the information bias on association study. Specifically, we designed our simulation studies based on the characteristics of the Electronic Medical Records and Genomics (eMERGE) Network. Through simulation studies, we quantified the loss of power due to misclassifications in case ascertainment and measurement errors in covariate status extraction, with respect to different levels of misclassification rates, disease prevalence, and covariate frequencies. These empirical findings can inform investigators for better understanding of the potential power loss due to misclassification and measurement errors under a variety of conditions in EHR based association studies.

Introduction

Along with the widely use of electronic health record (EHR) systems over the last decade, huge amounts of longitudinal patient information, including coded/structured data (e.g., the International Classification of Disease codes) and clinical narratives (e.g., as admission notes and discharge notes), have been accumulated and are available electronically. These large clinical databases are valuable data sources for clinical, epidemiologic, genomic, and translational research, where identifying patients with specific diseases is often the critical step. However, manual review of patients’ charts is extremely time-consuming and costly as the decision criteria are often complex and require review of different sources of information by domain experts. Many studies have shown that structured data such as the ICD codes are not sufficient enough to determine the patient cohorts as much of the detailed patient information was recorded in clinical narratives. Therefore, a large number of phenotyping methods have been developed to combine the clinical narratives with coded data to identify patients with certain diseases. Due to the sensitiveness of clinical data and intrinsic difference between diseases, current phenotyping studies are often performed in a disease-specific manner, where training and test corpus are developed for each disease. Many studies have focused on the prevalent diseases, such as diabetes, hypertension, and rheumatoid arthritis, or some common healthcare problems associated with significant morbidity and mortality, such as heart failure, colorectal cancer, and venous thromboembolism. Some studies also focus on observational characteristics such as smoking status, obesity. Various clinical NLP systems have been applied to extract clinical information from text to facilitate phenotype identification, e.g., MedLEE, MetaMap, KnowledgeMap and cTAKES. Some of them provide individual module to identify common phenotypes, e.g., the smoking status detection module in cTAKES. Both rule-based and machine learning based systems have been developed to determine phenotypes. Most of the rule-based methods heavily involve domain experts’ knowledge, and the machine learning based phenotype identification methods often achieve better performances by leveraging different sources of information. The performances of phenotyping methods vary greatly among different diseases, and perfect genotype identification is difficult to achieve.

Although EHR-based clinical observational studies have been successfully performed based on automated phenotyping solutions, there is no systematic study to examine the misclassification of outcome and/or measurement...
error in covariates caused by imperfect phenotyping algorithms. In epidemiologic studies designed for EHR data, researchers have already noticed that the statistical models suffered from the information bias caused by the misclassification and/or measurement errors. Many studies have demonstrated that naive analyses ignoring misclassification lead to biased results in a variety of settings. As a result, the last 15 years have seen a rise in statistical methods to accommodate misclassification of disease status and measurement error in covariates, specifically when estimating odds ratios (ORs) or relative risks (RRs) in binomial regression; unbiased estimation of these parameters is typically of interest in pharmacoepidemiologic research. Several statistical methods have been developed to correct for a misclassified outcome or to correct for a misclassified covariate, for example using matrix methods, inverse matrix methods and maximum likelihood methods. However, the loss of power due to the misclassification of outcomes and measurement error in covariates has not been fully understood.

In this study we perform a systematic analysis on the power loss of EHR-based association studies due to misclassified outcomes and measurement errors in covariates through simulation studies. We aim to discuss this problem in two general types of studies. In the settings of genetic association studies using EHR-based phenotypes, we focus on the impact of misclassified binary outcomes (i.e., misclassified phenotypes) only. In the settings of epidemiological association studies using EHR-based disease status and covariates, we investigate the further impact of measurement errors in covariates on the association test. Various settings of misclassification rates, magnitudes of measurement errors and levels of factors that would influence the power are considered. Simulation designs, parameter settings and models to be compared are introduced in the Methods section. Power curves and other simulation results are shown and discussed in the Simulation Results section. Further discussion on the power loss, type I error and effect size estimation are shown in the Discussion section. Our study highlights the issue of power loss in association test with the presence of measurement inaccuracy, which should be considered in all association studies that use data from the EHR systems.

Methods
Simulation setups
To investigate the impacts of misclassification of binary outcome (disease status) and measurement error in covariates on association studies, we start with the non-differential misclassification where the misclassification rates do not vary across covariate categories. Also, the measurement error of the covariate is assumed to be independent with the outcome. We consider two general types of studies: the genome-wide association studies (GWAS) where the outcome is EHR-based binary disease status and the covariate is the genotype at a single nucleotide polymorphism (SNP) locus, and the epidemiological studies where both the outcome and covariate are EHR-based (e.g., type II diabetes and smoking status). In this empirical study, we conduct simulations to mimic the setups of these two types of studies. We consider different scenarios by choosing different parameters of disease prevalence, covariate frequency, and misclassification rates in both outcome and covariate.

The prevalence of type II diabetes and multiple sclerosis are chosen to represent the common disease and the rare disease respectively. The prevalence for the common disease is 37.5% and for the rare disease is 3.2% according to recent literatures. To mimic the performance of some commonly used phenotyping algorithms, we choose the misclassification rates of the outcomes to be similar with the eMERGE algorithm of identifying phenotypes from EHR clinical texts. This algorithm has been used in many EHR related studies and its performance has been reported in different literatures. According to Ritchie et al (2010), in a relatively high-performance situation for phenotyping type II diabetes, the positive predicted value (PPV) of the algorithm is approximately 0.9 and the sensitivity is 0.84. The specificity is then calculated to be 0.96 as in Ritchie et al (2010). In a relatively low-performance situation, the sensitivity and specificity of the eMERGE algorithm are reported to be around 0.665 and 0.819, respectively. The misclassification rates of the phenotyping algorithm often depend on the prevalence of the disease. Phenotyping a rare disease tends to be more accurate than a common disease. For rare diseases (e.g., multiple sclerosis), the sensitivity of the eMERGE algorithm is around 0.857 and the specificity is 0.997 in a high-performance situation, and 0.707 and 0.988 in a low-performance situation. We fix the sample size at 5,000, and consider a sequence of effect sizes (in the scale of log odd ratio) in a logistic regression to evaluate the statistical power. These settings are the same for both genetic and epidemiological studies, and number of simulation is 1,000 for each setting.

In genetic association studies, the genetic information (i.e., genotypes) is often quite accurate with low genotyping errors. Therefore, we consider only the outcome misclassification in the GWAS settings. Minor allele frequency (MAF) of SNPs is set to be 0.2 for common variant scenarios and 0.03 for rare variant scenarios. In each scenario, two logistic regressions are fitted using the true outcomes and the misclassified outcomes, respectively. The reject rates are reported for each effect size, and power curves are plotted. The difference in power between using the true outcome and the misclassified outcome quantifies the impact of misclassification in each.
In almost all GWAS, a large number of SNPs are investigated. In order to understand the overall power loss for a typical genetic association study, averaged powers are obtained by sampling MAFs from a real genetic dataset (1000 Genomes Project Consortium, 2012) without replacement in each simulation to mimic the situation in a real GWAS.

In epidemiological studies, however, categorical covariates such as smoking status are often measured with different amount of measurement errors. Hence, we considered both the outcome misclassification and covariate measurement error in the epidemiological settings in order to evaluate the further impact of the measurement error in the covariate on association test. In some EHR systems, smoking status is recorded as a discrete variable with three categories (i.e. the classified smoking status): non-smoker, past-smoker, and current smoker. In this case, we use a single categorical variable which takes values 0, 1 and 2 to mimic the smoking status (non-smoker, past-smoker, and current smoker). The true status of the covariate is then generated from a multinomial distribution where the proportions of the three smoking status categories are based on the results in Liu et al (2012)\(^\text{17}\). The misclassified covariate is generated by mimicking the performance of the cTAKES smoking status detection module\(^\text{17}\). More precisely, as discussed in the paper, the algorithm first classifies subjects into non-smokers and smokers, and then classifies the smokers into past smokers and current smokers. Both classifiers involve misclassification and the misclassification rates are listed in Liu et al (2012)\(^\text{17}\). For scenarios with small amount of measurement errors (e.g., customized cTAKES module for eMERGE data), the sensitivity and specificity are approximately 0.98 and 0.97 for the first classifier and 0.94 and 0.92 for the second classifier. For scenarios with relatively large amount of measurement errors (e.g., direct use of cTAKE module without customization), the sensitivity and specificity are around 0.31 and 0.91 for the first classifier and 0.84 and 0.51 for the second classifier. In some other EHR systems, smoking status is recorded as the self-reported number of cigarettes per day (i.e. the quantified smoking status). To mimic this type of covariate data, we generate the smoking status as a variable taking non-negative integers using a zero-inflated Poisson distribution, considering a lot of patients are non-smokers. According to Kiviniemi et al (2011)\(^\text{3}\), the non-smoking proportion in the US is around 83% and the average number of cigarettes per day for one person is 15. The measurement error in smoking status is generated by randomly adding or subtracting a random error which follows a negative binomial distribution. For the situation with relatively small amount of measurement errors, we generated the random error from a negative binomial distribution with mean 0.5 and variance 0.625; and for large amount of errors, the random error are generated from a negative binomial distribution with mean 3 and variance 7.5. Figure 1 shows the scenarios what we consider and the corresponding parameter we choose in this paper.
In all scenarios, the smoking status is treated as ordinal variable assuming the effect of smoking on presence of disease is additive. Powers are calculated for different effect sizes from 0 to 1 (in the scale of log odds ratio) in the GWAS setting and the epidemiological setting with classified smoking status, and for different effect sizes from 0 to 0.05 in the epidemiological setting with quantified smoking status.

**Methods under comparison**

Under genetic settings, we compare the model using true disease status with the model using misclassified outcome. Under epidemiological settings, we compare the model using true disease status and true smoking status with the model using misclassified disease status and misclassified smoking status. The models to be compared are described below:

1. Logistic regression using true disease status and true covariate.

   Let \( Y \) denote the true disease status (e.g., type II diabetes) of a patient, and \( X \) denote the true covariate status (e.g., genotype at a SNP locus, or true smoking status), we assume a logistic regression model: 
   \[
   \text{logit} \left( \Pr(Y = 1) \right) = \beta_0 + \beta_1 X,
   \]
   where \( \Pr(Y = 1) \) is the probability of a patient having the disease.

2. Logistic regression using misclassified disease status and true covariate.

   In the genetic association studies using the EHR-based phenotypes, instead of observing the true disease status, we could only obtain the disease status of the patients with misclassification. Let \( Y^* \) denote the surrogate of the true disease (e.g., type II diabetes status extracted from EHR data), and \( X \) denote the genotype at a certain SNP. The logistic regression model using the surrogate as the true type II diabetes status could be expressed as: 
   \[
   \text{logit} \left( \Pr(Y = 1) \right) = \gamma_0 + \gamma_1 X.
   \]
   Here, the probability of having type II diabetes (\( \Pr(Y = 1) \)) is replaced by the probability of observing the patient in the disease category (\( \Pr(Y^* = 1) \)). This model is not correct since these two probabilities are not equal and: 
   \[
   \Pr(Y^* = 1|Y = 1) = \alpha_1, \quad \Pr(Y^* = 1|Y = 0) = 1 - \alpha_2,
   \]
   where \( \alpha_1 \) is known as the sensitivity and \( \alpha_2 \) is the specificity of the phenotyping algorithm.

3. Logistic regression using misclassified disease status and covariate with measurement error.

   In epidemiological studies using EHR-based phenotypes and covariate (e.g., type II diabetes and smoking status), both the outcome misclassification and covariate measurement error are considered. Let \( Y^* \) denote the surrogate of the true disease status and \( X^* \) denote the surrogate of true smoking status. The logistic model using surrogates of both disease status and smoking status could be expressed as: 
   \[
   \text{logit} \left( \Pr(Y = 1) \right) = \eta_0 + \eta_1 X^*.
   \]
   It also holds that \( \Pr(Y^* = 1|Y = 1) = \alpha_1, \quad \Pr(Y^* = 1|Y = 0) = 1 - \alpha_2 \). When the smoking status is recorded as non-smoker, past-smoker and current smoker, the misclassification of smoking status involves two classifiers and satisfies:
   \[
   \begin{align*}
   &\Pr(X^* = 1, 2|X = 1, 2) = \alpha_a, \quad \Pr(X^* = 1, 2|X = 0) = 1 - \alpha_b, \\
   &\Pr(X^* = 2|X = 2) = 1 - \alpha_c, \quad \Pr(X^* = 2|X = 1) = 1 - \alpha_d,
   \end{align*}
   \]
   where \( \alpha_a \) and \( \alpha_b \) are the sensitivity and specificity for the first classifier and \( \alpha_c \) and \( \alpha_d \) are the sensitivity and specificity for the second classifier of the smoking status detection algorithm.

In each setting, we are interested in testing for an association between the covariate (smoking status) and disease (type II diabetes) by testing the association parameter being 0, i.e., \( H_0: \beta_1 = 0, \gamma_1 = 0, \) or \( \eta_1 = 0 \) in the models.

In the simulation studies, we generate the true status of disease and covariate from model 1, and generate the surrogates using the assumed misclassification rates and error distribution. Powers are compared between the misclassified model and the true model in order to quantify the impacts of misclassifications of outcomes and measurement errors of covariates on association test.

**Simulation results**

1. Genetic settings.

   Figure 2 presented the power curves of the genetic association tests in settings with common and rare diseases. When true disease statuses were known, the power for detecting association was higher for common diseases compared to rare diseases, and higher for common variant compared to rare variant when controlling for other parameter values. Such a finding also held when true disease statuses were unknown and surrogates were used. This finding was consistent with the experiences in association studies\(^{15,16}\). To evaluate the impact of misclassifications, we compared the power of using true outcomes against using the surrogates within the same setting of disease prevalence and allele frequency. The left panel in Figure 2 suggested that for a common variant (i.e., high allele frequency), the loss of power was relatively small as long as the misclassification rates were low (the maximal loss of power is around 15%). On the other hand, if the variant was rare (i.e., low allele frequency), the loss of power due
to the misclassification was sizable (up to 20%) when misclassification rates were low. However, when misclassification rates were high, the losses of power were all larger than 50% for both common and rare variant scenarios. The right panel in Figure 2 suggested a similar finding, except the corresponding power was lower due to the fact that the disease was rarer.

Another observation from the solid and dashed curves in the left panel of Figure 2 was that when the true log odds ratio was 0.3 (i.e., odds ratio of 1.35), the association test based on the true outcome (or misclassified outcome with small misclassification rates) had a power of 100%. In other words, misclassification had no impact on loss of power, provided the effect size was moderately strong, both disease and covariate were common, and misclassification rates were relatively small. If any of these conditions failed to hold, the loss of power due to misclassification was still substantial. For example, for the same effect size, when considering the common disease and rare variant scenario, the power loss due to misclassification was 17.3% for small misclassification rates and 46.2% for high misclassification rates. These numbers were 10% and 29.6% respectively, for rare disease and common variant, and 2% and 7.2% for rare disease and rare variant. In the setting of no misclassification, the power for the rare disease and common variant scenario and the rare disease and rare variant scenario were 68.7% and 18.6%. Therefore the power loss could not be ignored.

Figure 3 presented the power for association tests at a single SNP in different settings. For a typical GWAS, millions of SNPs are tested where both rare variants and common variants are included. We are often more interested in the averaged loss of power due to misclassifications, where the loss of power in association test at each SNP is averaged over millions of SNPs. To obtain the averaged power, we sampled the allele frequencies without replacement from the data from the reference panels of The 1000 Genomes Project (here we used Utah Residents with Northern and Western European Ancestry, or CEU) (1000 Genomes Project Consortium, 2012). The empirical distribution of the minor allele frequencies was displayed in Figure 2. The SNPs with a MAF greater than 0.1 were distributed approximately evenly, and SNPs with MAF less than 0.05 were considered rare in this population.
Figure 4 showed the averaged power curves for a genetic association study based on the CEU population. Power curves in the left and right panels showed the impact of misclassified phenotypes on association testing for common and rare diseases. These power curves implied that from an overall perspective how much power was lost due to misclassified outcomes in different settings of misclassification rates. Since the minor allele frequency of most SNPs in the genetic association study was distributed from 0.1 to 0.5, the averaged power curves were similar as the common variant scenarios in Figure 3. It suggested that when misclassification rates were low, the power loss was relatively mild compared to the power loss when misclassification rates were high. For common diseases, when the testing allele had an effect size (log odds ratio) larger than 0.3 and under low misclassification rates, using the surrogates did not cause power loss greater than 1%. When misclassification rates were high, however, the log odds ratio had to be at least 0.5 to reach a power loss less than 5%. When diseases were rare, the log odds ratio had to be greater than 0.4 for the low misclassification scenario, and 0.6 for the high misclassification scenario to ensure a power loss less than 5%.

To evaluate the power loss in extreme situations, we calculated the maximal power loss when the effect size was varying from 0 to 1. When misclassification rates were low, the maximal power loss was 15% for common disease and 11% for rare disease. While when misclassification rates were high, the maximal power losses were 50% and 29% for common and rare diseases correspondingly. Therefore, when the sensitivity and specificity of the phenotyping algorithm were not high enough, the average power loss could be huge for SNPs with relatively moderate effect sizes for both rare and common diseases.

2. Epidemiological settings.

In this section, we mimicked the settings of epidemiological studies using EHR-based disease and covariate status (i.e. smoking status). Besides the misclassified binary outcomes, we evaluated the further impact of measurement errors in covariates on the power loss of association test. The misclassification settings for the disease status and smoking status could be found in Figure 1.

When the smoking status only contained three classes, Figure 5 showed the power curves in different settings of misclassification and measurement errors for common and rare diseases. The previous finding that the power for rare diseases is lower than common diseases when controlling for other settings also held in the epidemiological studies. The comparison between the solid line and the four dashed lines within each panel showed the impact of different levels of misclassification and measurement errors on the power of association testing.

Combined with Figure 3 and Figure 5, it suggested that adding the measurement error increased the power loss compared with models with only the outcome misclassification. For example, when the effect size (log odds ratio) was 0.3 for common disease, the power losses of the four scenarios in the plots (low misclassification small measurement error, high misclassification small measurement error, low misclassification large measurement error and high misclassification large measurement error) were 2.7%, 35.8%, 48.8% and 77.6%. For rare disease, the powers of the four scenarios were 15.1%, 30.6%, 45.8% and 48%, respectively. It also showed that only when both
misclassification and measurement error were small, the power loss was relatively small. When either outcome or exposure had a high misclassification rate, the power loss was severe.

Figure 5 Comparison of Power for common and rare diseases in epidemiological settings where the covariate contains three categories (non-smoker, past smoker and current smoker). Misclassified outcome* stands for high sensitivity and specificity (Sensitivity=0.84, specificity=0.96 for common diseases; sensitivity=0.857, specificity=0.997 for rare diseases). Misclassified outcome† stands for low sensitivity and specificity (Sensitivity=0.665, specificity=0.819 for common diseases; sensitivity=0.707, specificity=0.988 for rare diseases). Covariate* stands for small amount of covariate misclassification (sensitivity and specificity were chosen as 0.98 and 0.97 for the first classifier and 0.94 and 0.92 for the second classifier). Covariate† stands for large amount of covariate misclassification (the sensitivity and specificity were set to be 0.31 and 0.91 for the first classifier and 0.84 and 0.51 for the second classifier).

Figure 6 Comparison of Power for common and rare diseases in epidemiological settings where the covariate is the self-reported number of cigarettes per day. Misclassified outcome* stands for high sensitivity and specificity (Sensitivity=0.84, specificity=0.96 for common diseases; sensitivity=0.857, specificity=0.997 for rare diseases). Misclassified outcome† stands for low sensitivity and specificity (Sensitivity=0.665, specificity=0.819 for common diseases; sensitivity=0.707, specificity=0.988 for rare diseases). Covariate* stands for small amount of covariate measurement error in number of cigarettes which was generated from NB(0.5,0.625). Covariate† stands for large amount of covariate measurement error which was generated from NB (3, 7.5).

When using the number of cigarettes to code the smoking status, the power curves under different scenarios were showed in Figure 6. The major findings under this scenario were the same as the previous scenario in which the smoking status was recorded as three categories. When the outcome misclassification was low and measurement error of the covariate was small, the power loss was within 18.2% for common diseases and 12.4% for rare diseases.
However, when either the outcome misclassification or the covariate measurement error was large, the power loss was severe up to 50%.

Through this simulation study, we quantified the impact of misclassification and measurement error in association studies. It was showed that rare diseases, compared to common diseases, had lower power for testing the association controlling for the prevalence of covariate, misclassification rates and measurement error amount. The prevalence of the covariate (minor allele frequency in this study) also influenced the power in the sense that rare variant scenarios had lower power than the common variant scenarios when controlling for other settings. Therefore, in practical researches, study designs should account for these factors to adjust the sample size to achieve a proper testing power. When only outcomes were misclassified, the power loss was small when the misclassification rates were low, and it was sizable when misclassification rates were high. Moreover, when the covariate and the outcome were not measured accurately, the power loss increased compared to the models with only the outcome misclassification. In these scenarios, power loss was relatively large when either the measurement error or the outcome misclassification rates were large.

Discussion

In this empirical study, we conducted simulation studies under different values of misclassification rates and amount of measurement errors to evaluate the loss of power in both genetic and epidemiological association study settings. We concluded that factors including disease prevalence, covariate frequency, misclassification rates of the disease status as well as the amount of measurement errors all influenced the power of association test. The power loss was relatively small as long as the misclassification rates and/or measurement error were low. Otherwise, power loss could be substantial and was increased with higher misclassification rates and/or measurement error rates. From the simulation results, we also observed that the type I error of the association test that ignored misclassification and measurement errors were not inflated under all scenarios. In other words, although the outcome and covariate could be measured with errors, the naive logistic regression ignoring these errors can still control type I error well as long as the misclassification is nondifferential. This is consistent with the findings of Neuhaus (1999) and Li & Duan (1989). Therefore, the only issue of using the naive logistic regression is the power loss.

However, while the EHR data are widely used in all kinds of association studies, the problem of power loss due to inaccuracy of EHR-based phenotypes and covariate records are sometimes ignored. Our study investigated this problem in two major types of studies and showed that when observations were measured with certain amount of errors, the power loss could be quite substantial. If no adjustment is made to account for the misclassification and measurement errors, sample size should be recalculated to deal with the loss of power.

In practical studies, when sample size is fixed and misclassification exists, association tests that account for the misclassification are needed. Moreover, the estimations of the effect sizes can also be biased and need to be corrected. Neuhaus (1999) and Li & Duan (1989) pointed out that when the outcome and covariate were misclassified, the estimated effect size (log odds ratio) is biased towards the null if using naive logistic regression. One limitation of our empirical study is that we only considered two levels of disease prevalence, minor allele frequency, misclassification rates and measurement error amount. The scope of scenarios that we have discussed in this paper is still limited. For example, in real applications, the disease prevalence could be lower than 3.2% as we considered here, and misclassification rates could vary from a wider range. However, the results of this study are useful to inform and guide investigators on the magnitude of loss of power in certain scenarios. And the conclusions about the relationships between powers and each factor could help the investigators obtain a rough estimation of potential loss of power in their studies.

In this study, we considered a relatively simple situation where the misclassification of the outcome is independent of the covariate, as well as the measurement error of the covariates. This assumption might not hold in some practical situations since the error probabilities might differ across different covariate levels. Moreover, the amount of measurement error of the covariate might be correlated with the disease misclassification rates. Empirical evaluation of power loss in these more complicated situations will be reported in the near future.

Furthermore, as we concluded that the power of the association test is influenced by various factors such as the disease prevalence, frequency of the covariate, as well as the misclassification rates, it is of interest to investigate how the power loss is attributable to these factors individually and jointly. Our future work also includes the extension from simple univariate logistic regression to multivariate regression to better understand the impact of misclassification on the power loss of association test in the presence of confounders.

Conclusion

Over the last decade, EHR systems have been increasingly implemented at US hospitals. Substantial amounts of
detailed patient information, including lab tests, medications, disease status, and treatment outcome, have been accumulated and are available electronically. These large clinical databases are valuable sources for clinical and translational research. Despite their great potential, the complex and uneven nature of clinical documentation and data quality brings additional challenges for analyzing EHR data. A critical challenge is the information bias due to the measurement errors in outcome and/or covariates. We conducted empirical studies to quantify the impacts of measurement errors on EHR based associated study. More specifically, we designed our simulation studies based on the disease prevalence, covariate frequencies and misclassification rates using data from the Electronic Medical Records and Genomics (eMERGE) Network. Through simulation studies, we quantified the loss of power due to misclassifications in case ascertainment with respect to different levels of misclassification rates (e.g., an algorithm with high precision and recall vs one with low precision and recall), disease prevalence (e.g., common disease vs rare disease), and covariate frequencies (e.g., common allele frequency vs rare allele frequency). We also evaluated the further loss of power if the covariate conditions were subjected to measurement errors, such as smoking status extracted from medical records by a certain algorithm. These empirical findings can help to inform investigators for better understanding the potential power loss due to different types of measurement errors under a variety of conditions in EHR based association studies.

Acknowledgement
This study was supported by grants from the NLM 2R01LM010681-05, NIGMS 1R01GM103859, 1R01GM102282, CPRIT R1307, R21 LM 012197, and AHRQ R03HS022900. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

References


Supporting Common Ground Development in the Operation Room through Information Display Systems

Yuanyuan Feng, MS, and Helena M. Mentis, MS PHD
1University of Maryland Baltimore County, Baltimore, MD, USA

Abstract

Effective information sharing is crucial for clinical team coordination. Most information display systems have been designed to replace verbal communication. However, information may not be available for capture before a communication event and information needs often become clear and evident through an evolving discourse. Thus, to build tools to support clinical team in situ information sharing, we need a better understanding of how evolving information needs are identified and satisfied. In this study, we used sequential analysis techniques to explore the ways communication and information sharing events between an attending surgeon and a resident change throughout a laparoscopic surgery. We demonstrate how common ground is developed and maintained, and how information needs change through the efforts of grounding. From our findings, we suggest that the design for information display systems could encourage communication and support the articulation work that is necessary to accomplish the information sharing.

Introduction

Timely and accurate information sharing is crucial for the coordination of clinical team members in dynamic and information-rich clinical environment1. Failures in communication have been shown to be a major contributor to medical errors and adverse events in healthcare2. Because of these collaborative challenges, there has been extensive effort in designing systems to replace verbal communications, such as hand-off tools, by retrieving, integrating and visualizing information to the team3-6. For instance, in the study of team communication in the cardiac operating room, Parush et al. proposed an augmented display to increase the team situation awareness of patient information, vital signs, and critical events5. However, the underlying assumptions in this paper and others like it has been that (1) the information is already stored somewhere in the system, and (2) health care professionals are aware of their information needs. These assumptions are in contrast to studies that highlight how communication contains a high volume of information that may not be collected before a communication event and, through an evolving discourse, information needs become clear and evident.

For instance, Coiera highlights how communication and information tasks take place on a continuing basis where informal tasks are accomplished with discourse-embedded information7. This practice is manifest in the survey conducted by McKnight et al., which showed that clinicians preferred informal communication instead of the time-consuming explicit information seeking process8. Furthermore, in the study of collaborative information behavior in the clinical settings, Reddy et al. had shown that communication not only allows team members to identify and share information, but also ties together different pieces of information to make sense of the context9. Moreover, information needs have been found to be highly complex, which requires team members to clarify individual information seeking tasks and to integrate their tasks together to generate knowledge10.

The concept of common ground can further elucidate how effective information sharing between team members is embedded in a context and requires ongoing effort16. A lack of previous collaborative activity leads to a large hurdle for the team to quickly and effectively share the relevant content and coordinate the work process; not least in part due to the team members’ different levels of expertise17. This increases the need for team members to quickly build enough shared knowledge. In particular, they must establish that they both engender the same knowledge of how to perform the task as well as, throughout the collaboration, maintain that they both have a shared knowledge of the task domain18. Common ground entails an ongoing process of checking, clarifying, and sharing further knowledge in order to maintain this shared understanding19. This process of maintaining common ground actually enables greater efficiency or minimal effort for communication over time for the team20. Studies of teamwork in critical settings, such as flight crews, have shown that more efficient communication among team members results in better coordination and performance. This is especially relevant for supporting teams of experts with diverse skillsets collaborating on critical problems with time constraints21.
To build tools that support clinical team collaboration, we need a better understanding of the mechanisms through which common ground is developed and maintained. Although previous research has demonstrated that common ground increases through joint experience on a task over time, scrutinizing the grounding process of a clinical team in an information-rich environment leads us to ask a series of questions regarding information needs. For example, how does the information need change over the course of team collaboration? How is the information need identified and satisfied? The answers to these questions are essential for designing systems to display information and support collaborative activities. By identifying the grounding process, we can make informed decisions on when and how to provide appropriate tools.

One information-rich medical environment that has garnered recent attention in the design of information collection and display systems is surgery. With the increasing rise of minimally invasive surgery, surgeons are required to coordinate around highly specialized medical images such as fluoroscopic x-rays and laparoscopic video throughout a surgery. In most cases, surgeons rely on efficient and effective communication practices to successfully coordinate with each other. This is all the more complex when, as in teaching hospitals, one of the coordinating surgeons has a reduced level of knowledge as in the case of a resident. Many residents practice their surgical skills and learn the process of surgical procedures through assisting in surgeries. Although a resident may have performed a particular surgery before, they are certainly not yet proficient. Oftentimes, the resident and attending surgeon have not worked together before or, if they have, it is not the specific surgery they are about to perform. Thus, not only do they have to coordinate their actions, but the attending surgeon also has to assess the resident’s level of knowledge of the case and determine how much direction they need to perform the procedure.

In this paper, we present a conversation analysis study of in situ laparoscopic surgical training to show the considerable information generated and knowledge gained throughout the discourse. We used sequential analysis techniques to explore the ways communication is conducted and demonstrate how common ground is developed and maintained over the course of a laparoscopic surgery. Examining the communication patterns exposes not only how attending surgeons pass on their knowledge and skills to residents through verbal communication, but also how information needs change through the efforts of grounding. From our findings, we address the importance of communication patterns in information acquisition and sharing, as well as question the predominant direction for information display systems in the operating room.

**Study Design and Data Collection**

The study was based on videos recorded during fieldwork in the operating rooms of the Cambridge Hospital. Human Subject’s approval was granted through the Cambridge Health Alliance IRB.

A total of ten laparoscopic cholecystectomy cases were video recorded. The cases included three different surgeons and three different PGY4 residents in different combinations of the two groups (Table 1). In each case, the surgery was performed primarily by a senior resident with assistance from an attending surgeon. The videos were transformed into picture-in-picture format, including the external video showing the interaction between attending surgeons and residents, along with the internal video captured by the laparoscopic video recorder. The communication was recorded with a wireless microphone worn by the attending surgeon.

**Table 1.** Arrangement of instructor and resident among the ten cases.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Attending Surgeon</th>
<th>Resident</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>c</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>b</td>
<td>1</td>
</tr>
</tbody>
</table>

**Coding Scheme and Procedure**

In our cases, the surgeons were collocated and verbally communicated over the laparoscopic video, which displayed the operative field. In our analysis, we focused on the utterances made by the surgeons. The communication content was coded manually based on a dialogue act scheme, which was developed to
understand the development of common ground among interdependent team members in emergency management tasks to provide implications for the design of collaborative systems24 (Table 2). This method emphasizes the use of communicative intentions to interpret the development of common ground. It includes two types of common ground – content and process. Content common ground is defined in the coding scheme as “the shared understanding on the subject and focus of work”, while process common ground is defined as “the shared understanding of the rules, procedures, timing and manner in which the interaction will be conducted”25.

We used this coding scheme to understand the processes of knowledge sharing and common ground building in OR-based training for laparoscopic surgery. Specifically, dialogue acts provided us insights into the ways that information was gained, the kinds of information that were transferred, and the collaboration between attending surgeons and residents.

Table 2. Dialogue act codes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Dialogue Act</th>
<th>Description</th>
<th>Utterance Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer Info</td>
<td>Add Info (AI)</td>
<td>Provides new information, not elicited.</td>
<td>“The gallbladder is going to be very thin walled.”</td>
</tr>
<tr>
<td></td>
<td>Query (Q)</td>
<td>Question used to elicit new information.</td>
<td>“What’s that?”</td>
</tr>
<tr>
<td></td>
<td>Reply (R)</td>
<td>Reply to query to provide new information.</td>
<td>“Yeah there is probably a peritonaeum.”</td>
</tr>
<tr>
<td>Check Understanding</td>
<td>Check (CH)</td>
<td>Verify own understanding of information previously presented by others.</td>
<td>“So, we’re gonna be in and out of here?”</td>
</tr>
<tr>
<td></td>
<td>Align (AL)</td>
<td>Verify partner’s understanding of information previously presented to others.</td>
<td>“So you know I like to make an imaginary line right below this.”</td>
</tr>
<tr>
<td></td>
<td>Clarify (CL)</td>
<td>Clarifies or restates information already presented.</td>
<td>“So, that’s the common duct.”</td>
</tr>
<tr>
<td></td>
<td>Acknowledge (AC)</td>
<td>Signals receipt of information, understanding.</td>
<td>“Alright.”</td>
</tr>
<tr>
<td>Manage Process &amp; Decision</td>
<td>Manage (MN)</td>
<td>Instruction, command, direct or indirect request for action; orchestrating strategy, how to do the work</td>
<td>“I’m going to hold this side, you take yours.”</td>
</tr>
<tr>
<td></td>
<td>Summarize (SA)</td>
<td>Summarizes information previously presented.</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Judge (J)</td>
<td>Individual judgment, opinion, or preference.</td>
<td>“The cystic was extremely short.”</td>
</tr>
<tr>
<td></td>
<td>Confirm (CO)</td>
<td>Requests partners’ agreement on a proposed decision.</td>
<td>“For the top and the bottom? Or just right under the stitch?”</td>
</tr>
<tr>
<td></td>
<td>Agree (AG)</td>
<td>Indicates approval for a prior judgment or decision.</td>
<td>“Yeah just put that down.”</td>
</tr>
</tbody>
</table>

All utterances between attending surgeons and residents were coded using the dialogue act coding scheme by two coders. For better reliability, the coders viewed the videos and coded the transcripts independently. After the first independent coding session, the inter-rater reliability for the coders was found be to Kappa = 0.64. The two coders negotiated for any conflicting codes with one of the researchers who developed the coding scheme (last author of this paper). After this discussion, the two coders independently coded again and achieved an agreement of Kappa = 0.87. This was deemed high agreement26 and so the remainder of the cases were coded. The inter-rater reliability was analyzed using the Kappa statistic for each case. For any case with Kappa less than 0.7, the two coders negotiated the difference to achieve higher agreement. In the end, the average Kappa score is 0.78, ranging from 0.71 to 0.87.

Phase Identification

A well-established hierarchical task analysis of laparoscopic cholecystectomy identified five high-level surgical steps27. For our study, we used these surgical steps to define the five main phases of our laparoscopic cholecystectomy: patient preparation, gallbladder isolation, gallbladder removal, and cleanup. By using the hierarchical task analysis, we identified the start and end times of each phase in each surgery. The communicative events were then associated with a phase of the surgery by matching the communication occurrences to the phase time stamps.
Data Analysis

The focus of data analysis in our study was to identify the significant patterns of communication between an attending surgeon and a resident across phases in a laparoscopic cholecystectomy. We used lag-sequential analysis with log linear modeling to identify the sequential nature in the data. Since the lag sequential analysis emphasizes the consecutive nature of the data, it allows us to understand the immediate reactions to a given utterance, and further to identify the information needs for developing common ground. The non-parametric Friedman test was used to compare frequencies of significant communication patterns across phases. The data analysis was conducted in SPSS v 22.0 (Chicago, IL). A value of $P < 0.05$ was considered to indicate statistical significance.

Results

Of the ten laparoscopic cholecystectomy cases in this study, 657 minutes of operative time were assessed. During that time, a total of 4287 utterances were observed, among which 3221 utterances were between attending surgeons (87.6%) and residents (12.4%) with a median of 227.5 and a range from 126 to 910.

Lag-sequential Analysis

In order to identify the speakers of utterances, we labeled each utterance with the speaker as well – attending surgeon (att) or resident (res). Since we were interested in how the information was gained, we included repeated codes in our analysis. Thus, in total, we received 3161 two-chain communication events across 21 dialogue codes. Table 3 shows the average time spent and the frequency of communication events in each phase.

Table 3. Distribution of communication events across phases of a laparoscopic cholecystectomy

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
<th>Time Avg. (min)</th>
<th>Frequency of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patient preparation</td>
<td>8.1</td>
<td>330</td>
</tr>
<tr>
<td>2</td>
<td>Gallbladder isolation</td>
<td>27.7</td>
<td>1715</td>
</tr>
<tr>
<td>3</td>
<td>Gallbladder removal</td>
<td>18.0</td>
<td>894</td>
</tr>
<tr>
<td>4</td>
<td>Cleanup</td>
<td>11.9</td>
<td>221</td>
</tr>
</tbody>
</table>

We first constructed an initial three-dimensional matrix of Lag0 Dialogue Act (21) × Lag1 Dialogue Act (21) × Phase (4) to represent the sequential frequencies between dialogue acts across phases. Lag0 represents the initial utterance provided by either attending surgeons or residents. Lag1 presents the feedback provided by any surgeon given Lag0. The third dimension, Phase, serves as conditions, indicating the phase when the two-event chain occurred.

We fit the initial matrix into the log-linear model of complete independence (Lag0 + Lag1 + Phase). The test results indicate that the three variables are not mutually independent ($G^2 (1720) = 1964.627, p < 0.0001$). However, the test of conditional independence, including all main effects and all two-way interactions, fails to reject the null hypothesis that the sequential structures are independent of the phases given our initial matrix ($G^2 (1200) = 359.230, p = 1$). The results are not surprising, since there is sparseness in our high dimensional contingency table. In order to identify significant communication patterns in each phase, we split the initial matrix into four two-way contingency tables based on the phases and selected the partition that consisted of the rows and columns where both Lag0 and Lag1 were categories that had marginal frequencies greater than 5.

Table 4. Phase matrix fitted in the log linear model.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Observations</th>
<th>$G^2$</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>286</td>
<td>186.852</td>
<td>72</td>
<td>$p &lt; 0.0001$</td>
</tr>
<tr>
<td>2</td>
<td>1683</td>
<td>651.628</td>
<td>144</td>
<td>$p &lt; 0.0001$</td>
</tr>
<tr>
<td>3</td>
<td>856</td>
<td>381.787</td>
<td>144</td>
<td>$p &lt; 0.0001$</td>
</tr>
<tr>
<td>4</td>
<td>179</td>
<td>88.503</td>
<td>25</td>
<td>$p &lt; 0.0001$</td>
</tr>
</tbody>
</table>
After applying the cutoff value, testing the partition of the original tables revealed a highly significant interaction between Lag0 and Lag1 in all four phases (Table 4). The results indicate that there are significant patterns in each of the four phases.

**Common Communication Patterns Shared among Phases**

We proceeded to identify the significant patterns, which are defined by the value of adjusted residual greater than 1.96\(^2\). There are seven patterns shared by two or more phases (Table 5). As shown in Table 5, the shared communication patterns include both content information sharing and process management.

**Information Transfer**

Knowledge for the task content was either directly added or elicited by queries. The consecutively repeated adding information (AI) by the attending surgeons is the common pattern in all four phases. The high frequency of this pattern in Phase 2 indicates that a vast amount of information is explicitly presented in order to allow the residents to develop a mutual understanding with the attending surgeons. A non-parametric Friedman test of differences among repeated measures was conducted to compare the pattern frequencies across phases and rendered a Chi-square value of 13.194, which indicates that the frequency of the adding information pattern varies over the phases (p = 0.004). The post-hoc comparisons identified that Phase 2 has the most frequent patterns of adding information (p = 0.028), whereas the other three phases have similar frequencies to each other.

In the meantime, we also observed the residents’ efforts in adding relevant information (AI) in Phase 1, 2 and 3. This pattern reflects that the residents understood the content and were able to provide new information to supplement their mutual understanding. It indicates an increase in the common ground that allows residents to make an effective introduction of information.

**Table 5.** Observed frequencies, conditional probabilities, and adjusted residuals of common significant patterns across phases.

<table>
<thead>
<tr>
<th>Communication Pattern</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
<th>Phase 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att_AI - Att_AI</td>
<td>82</td>
<td>403</td>
<td>194</td>
<td>38</td>
</tr>
<tr>
<td>Att_AI - Res_AI</td>
<td>51</td>
<td>27</td>
<td>6.8%</td>
<td>7.6%</td>
</tr>
<tr>
<td>Att_AL - Res_AC</td>
<td>5</td>
<td>2</td>
<td>26.3%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Att_MN - Att_MN</td>
<td>30</td>
<td>274</td>
<td>170</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Communication Pattern</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
<th>Phase 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att_AI - Att_AI</td>
<td>53.9%</td>
<td>53.7%</td>
<td>54.5%</td>
<td>52.1%</td>
</tr>
<tr>
<td>Att_AI - Res_AI</td>
<td>(2.133)</td>
<td>(6.696)</td>
<td>(6.403)</td>
<td>(2.027)</td>
</tr>
<tr>
<td>Att_AL - Res_AC</td>
<td>6.8%</td>
<td>7.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Att_MN - Att_MN</td>
<td>39.0%</td>
<td>47.8%</td>
<td>52.8%</td>
<td>49.3%</td>
</tr>
<tr>
<td></td>
<td>(2.133)</td>
<td>(6.696)</td>
<td>(6.403)</td>
<td>(2.027)</td>
</tr>
</tbody>
</table>
Compared to adding information, the pattern of query (Q) and reply (R) is less frequent over the course of surgery. However, we observed significance in the two patterns of query and reply. The direct reply from the residents to the attending surgeons’ query among all four phases indicates that residents were able to understand a certain amount of queries and answer it in the next turn. By providing a reply, residents gained information through the process of recognizing the query, mapping it to the content, and relating it to the context. With the residents’ reply, attending surgeons assessed if more information needed to be provided to maintain the common ground. On the contrary, the residents’ request of information in Phase 1, 2 and 3 revealed the extra effort in gaining the necessary information. The residents could not proceed to perform a task without necessary information to develop the common ground; they needed to expend extra effort in formulating a request to gain the information. With the low frequencies of Q and R patterns, we have limited power to identify any significant increases in any phases.

Checking Understanding

Another significant pattern found in Phase 2 and 3 pertains to the alignment (AL) and acknowledgement (AC) acts. Without any feedback from the residents, the attending surgeons have to repeat the presented information to request residents’ feedback. Residents often replied with acknowledgement, showing that they had already known the information. The high frequency of attending surgeons’ alignment indicates that the residents failed to provide timely feedback and often received redundant information from the attending surgeons.

Process Management

To coordinate the procedure, attending surgeons often provided repeated explicit instructions (MN) over the course of surgery. The significant occurrence of this pattern indicates that the procedure was frequently coordinated and updated. The increased frequency of this pattern in Phase 2 and 3 shows that the attending surgeons provided detailed instructions to guide the residents step by step in the intricacies of that particular surgical procedure as Phase 1 and 4 contain knowledge that are more common across surgeries ($\chi^2 (3) = 17.022, p = 0.001$).

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Att.AI → Att.AG</th>
<th>Observed Frequency</th>
<th>Conditional Probability</th>
<th>Adjusted Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 2</td>
<td>Att.MN → Att.AG</td>
<td>15</td>
<td>26.2%</td>
<td>2.564</td>
</tr>
<tr>
<td></td>
<td>Res.Q → Att.AG</td>
<td>22</td>
<td>38.4%</td>
<td>2</td>
</tr>
<tr>
<td>Phase 3</td>
<td>Att.MN → Res.CH</td>
<td>7</td>
<td>21.8%</td>
<td>2.126</td>
</tr>
<tr>
<td>Phase 4</td>
<td>Att.MN → Res.CH</td>
<td>6</td>
<td>82.2%</td>
<td>3.002</td>
</tr>
</tbody>
</table>

Figure 1. Unique significant patterns in each phase.

Unique Communication Patterns in Each Phase

In addition to the common patterns, we identified 23 unique significant patterns in the four phases. Here we present the ones with higher observed frequency (Figure 1). The pattern in Phase 1 (patient preparation) shows that there is a significant connection between the attending adding information (AI) and then agreement (AG) with the course of the surgery (a process common ground dialogue act). This pattern reflects that after receiving relevant information from the attending surgeons, residents can make the right decisions for the process of the surgery. It indicates that with the mutual understanding of the content, the process common ground is easily achieved in Phase 1. In Phase 2 (gallbladder isolation), we observed
frequent patterns of attending management (MN) followed by attending agreement (AG) or resident query (Q). It shows that the attention of both attending surgeons and residents is focused on the process – attending surgeons provided explicit instructions and timely positive feedback to support residents moving forward, while residents follow the instructions and requests for further information to ensure that they share a mutual understanding with the attending surgeons. With increased content common ground, in Phase 3 (gallbladder removal), the communication pattern has changed from the residents requesting information to verifying their understanding of instructions (CH). This change indicates that residents have already developed the content common ground with regards to this particular surgery and have thus shifted their focus to the development of process common ground. In Phase 4 (cleanup), residents provide positive feedback to the instructions, indicating that they have achieved the mutual understanding of the process with the attending surgeons.

**Discussion**

Our study has aimed to identify the communication patterns and articulate the process of common ground development between a surgical attending and resident over the course of a laparoscopic surgery. These communication patterns reaffirm that information required to coordinate is embedded in the course of communication: through verbally sharing the information, surgeons and their trainees develop, maintain and update their common ground.

Our results illustrate different patterns between the development of content common ground and process common ground across phases. At the beginning of the surgery, both the attending surgeons and residents focused on developing the content common ground. As more critical moments occurred, they started to shift their attention to manage the process. This shift reflects the changes in residents’ information needs. In the initial phase, the adding of content information may help residents have a comprehensive operative view and become familiar with the anatomic structure. During the critical phases, such as Phase 2 and 3, residents began to make decisions about the process using information gained from previous phases. At these moments adding additional information may interrupt the residents’ thinking process and negatively influence their judgment and decision-making.

Based on the grounding process, we now discuss that how information needs emerge, how the information need is identified, and the cooperative nature of the communication. We identified that information needs emerge in three situations. First, the information need emerges when team members develop or update their mutual understanding of what is occurring during the different phases of the surgery. For instance, in the beginning of the surgery, there are a high number of communicative events regarding content between attending surgeons and residents – they needed to establish that they both engendered the same knowledge of the anatomic structures they were working on and their plans to perform the surgery. Secondly, as the surgery progresses the information need emerges when there is a need to understand a team member’s statements or decisions. There were situations when the residents could not understand the instructions. Instead of following the instructions, they verified their understanding or requested explanations from the attending surgeon. Finally, the information need is embedded in the information provided through discourse. This is reflected by the high frequency of consecutive acts of adding information by attending surgeons. After providing a piece of information, the attending surgeons realized there was other information that might support the residents’ understanding. Thus, we argue that the information needs are updated throughout the discourse, and communication is not merely an information transaction method, but a mechanism that generates information itself.

When there is an information need, the identification of this need is necessary for a communication event to be triggered. In this study, we found three ways for information needs to be identified. First, information needs can be identified through verbal requests. The verbal requests entail three different dialogue acts – query (Q), check (CH) and confirm (CO). The residents were found to commonly request new information from the attending surgeons, but they seldom confirmed their plans. This may be due to the high cognitive load in making plans compared to the limited efforts in accepting information. Secondly, information needs can be identified by the providers themselves, based on the mutual understanding between providers and receptors. For instance, the attending surgeons often elaborated on the typical instructions given to residents based on the inference that more information was needed. This direct act of adding information requires a mutual understanding developed between team members. Further, the information needs can also be identified through observing the actions. For example, after giving an instruction, the attending surgeons
identified the resident’s information need by observing how the residents executed this instruction. They would provide further instruction or add new information necessary for residents to accomplish the task. In this situation, the information providers are required to capture and understand their teammates’ action.

As we discuss above, communication is not only a way to transfer information, but also work that team members do to share and generate information – and this communication is a form of cooperative work. As is shown in the study, the attending surgeons provided instructions based on what their residents did or requested. The residents replied back based on the attending surgeons’ explanations of the instruction. In other words, they were mutually dependent in their efforts in communicating. Thus, the surgeons engaged in cooperative work of sharing information. If we regard the communication between surgeons as cooperative work, where multiple surgeons share and generate information, the articulation work that is required for communication is the work that supports the information accurately conveyed to and understood by the team. In this study, the articulation work entails knowledge level assessment, concept explanations, and feedback anticipation. For example, the attending surgeons asked questions at the beginning of the surgery to assess the residents’ knowledge levels, which informed them what kind of information to provide to the residents. When giving an instruction, the attending surgeons provided background knowledge or information about the operative field, in order for the residents to understand the content of the instruction and the context around the instruction.

Implications for Design of Information Display System

The understanding of the mechanisms through which common ground is developed and information needs change let us question the predominant direction for information display systems, which focuses on replacing verbal communications by directly retrieving and presenting the information to the team. Instead, the broad implication in our study is that the necessary information and the dynamic information needs are embedded in the communication between team members in the information-rich environment. For efficient communication it is not enough to simply allow surgeons to have more access to different kinds of information, but rather, timely mutual understanding of what content is presented, what is needed to be performed, and what should be verified. Thus, the direction for designing such systems should focus on encouraging efficient communication, as well as supporting the articulation work that is necessary to accomplish the communication.

This relates to our ongoing interests in developing interactive instructional systems that allow surgeons to add visual cues on the medical images to increase the efficient understanding of the utterances. In our previous work, we have identified the considerable work attending surgeons engage in in order to guide residents not only to capture but also to see anatomy within the videos in a laparoscopic surgery. In order to guide the residents’ gaze, the attending surgeons were found to put down the instrument and point at the video screen behind them, or to stop the task at hand and use the instrument to point at the structure, often coupled with an explanation. As indicated in this study, the pointing actions are articulation work required to convey the message. In order to support this work, we are designing a system that allows surgeons to sketch over or point at specific areas on the video in a touchless manner. This paper’s results show how the communication patterns, i.e. information transfer needs, change over the course of the surgery. We can use this knowledge to better understand which interaction mechanism is useful at which point in the surgery. For instance, according to our results, this system can be used by attending surgeons in adding information. Namely, adding information regarding content (i.e. anatomy) early in the surgery and then adding information regarding process (i.e. actions and locations to cut or move) during later phases. These engender different mechanisms of visualizing the information as well – i.e. mechanisms to circle anatomy for referring to anatomy versus mechanisms for pointing and gesturing direction of movement for process-related directions.

Further, we can use this knowledge to hypothesize what new visual interaction mechanisms can then actually help the communication during different phases. For example, we can support feedback anticipation by showing the residents’ understanding on the video. Eye gaze patterns can effectively provide point of attention to the partners. It has been shown that by indicating users’ attention allocation, eye gaze can improve the communication efficiency over visual representations. Eye gaze mechanism can be used in our system to visualize where residents are looking. Through identifying residents’ gaze variations, attending surgeons can provide instructions not only to train the eyes, but also to guide the residents in making decisions.
Limitations and Future Work

Our study is limited to laparoscopic cholecystectomy cases in the surgery department of one teaching hospital. Our results need to be further validated to be generalizable to communication patterns in different clinical settings. Residents also had a limited number of utterances. In our cases, residents manipulated the surgical instruments and performed major tasks, while attending surgeons served as the assistants retracting the gallbladder. In this situation, the attending surgeons needed to provide verbal instructions in order to lead the surgery. The residents, who were not as proficient, followed the instructions and performed surgical tasks. These practices elevated the residents’ cognitive load, which inhibited them to freely formulate utterances. Further, we only targeted the communication between attending surgeons and residents, without further exploring the process among all team members, including nurses, anesthetists and technicians. Other team members play essential roles in the successful coordination of the whole team. The communication patterns among all team members may be different from our results.

Our study shows that analyzing communication patterns based on the dialogue acts help us identify the change of information needs among clinical team members. In our future works, we plan to use this method to understand and evaluate new technologies deployed to support communication in the operating room. Any changes of communication patterns and information needs before and after the deployment may reveal the effectiveness of these technologies.

Conclusion

Communication is important for effective team cooperation. Our study shows the considerable amount of information generated in conversations over the course of a laparoscopic surgery and describes the process of common ground development between attending surgeons and residents. By scrutinizing the grounding process, we showed that information required to coordinate the process of surgery is embedded in the course of communication, and that the residents’ information needs varied across phases of surgery. Our results indicate that the design of information display systems in the operating room should not only encourage communication between team members, but also support the dynamic information needs that emerge through the efforts of grounding.

Acknowledgements

The authors would like to thank Mr. Christopher Wong for his help in coding the utterances. This research is sponsored by NSF Grant IIS-1422671.

References

33. Prokofieva A, Hakkani-Tur D, Slaney M. Eye gaze for understanding conversational speech. Spoken Language Technology Workshop (SLT); 2014: IEEE.
Exploring Dynamic Risk Prediction for Dialysis Patients

Malte Ganssauge, BS1, Rema Padman, PhD2, Pradip Teredesai, MD3, Ameet Karambelkar, MD3
1Technical University of Munich, Munich, Germany; 2Carnegie Mellon University, Pittsburgh, PA, USA; 3TMA, Pittsburgh, USA

Abstract

Despite substantial advances in the treatment of end-stage renal disease, mortality of hemodialysis patients remains high. Several models exist that predict mortality for this population and identify patients at risk. However, they mostly focus on patients at a particular stage of dialysis treatment, such as start of dialysis, and only use the most recent patient data. Generalization of such models for predictions in later periods can be challenging since disease characteristics change over time and the evolution of biomarkers is not adequately incorporated. In this research, we explore dynamic methods which allow updates of initial predictions when patients progress in time and new data is observed. We compare a Dynamic Bayesian Network (DBN) to regularized logistic regression models and a Cox model with landmarking. Our preliminary results indicate that the DBN achieves satisfactory performance for short term prediction horizons, but needs further refinement and parameter tuning for longer horizons.

1. Introduction

As the ninth leading cause of death in the United States, with prevalence exceeding 10% in the general population, Chronic Kidney Disease (CKD) is one of the big health burdens of the 21st century1–3. Patients on hemodialysis are in the end stage of chronic kidney disease (ESRD) and rely on dialysis treatments to survive. Despite significant advances in treatment over recent decades, patients still face severe risks and are in a vulnerable health condition. Approximately 24% of patients die within the first year on hemodialysis and hospitalizations are frequent4. The identification and early mitigation of individual risk is a challenging task for physicians who often make treatment decisions under severe time and cost constraints. Robust models that support the assessment of risk of adverse events and outcomes can help to identify early warning signals of increased risk, enable clinicians to target high risk patients, and can be used in the discussion of future treatment options with patients. A variety of prediction models have been suggested in the literature, including simple risk scores, classical regression models and machine learning methods5–12. However, they mostly focus on predictions for patients starting dialysis and are static with respect to time on dialysis or the changes in covariates. In this study, we explore dynamic models that allow updating a prediction when an individual’s health condition changes over time.

With the increasing use of electronic health records (EHR) for patient management, detailed individual information being documented over time has grown rapidly13. Leveraging this highly granular, longitudinal data for both differentiation of patients by disease progression stages and the integration of time-dependent covariates into models is challenging, but offer great potential to yield more accurate prediction estimates. In particular for chronic diseases, where individual data is often collected multiple times per year, the integration of such longitudinal observations is appealing. In this research, we focus on dynamic models that allow updating a prediction when an individual progresses in time. More specifically, our goal is to calculate for an individual patient $i$ at time $s$ the probability that the survival time $T_{\text{surv},i}$ is longer than a fixed horizon $w$, with the knowledge that the individual has survived until $s$ and we have observed some time-dependent covariates, $X_t(s)$:

$$P(T_{\text{surv},i} > s + w|T_{\text{surv},i} > s, X_t(s)).$$

We concentrate on the first year of hemodialysis where the highest mortality rates are observed and predict one-year survival ($w = 12$) at the start of dialysis and the end of each subsequent month ($s \in \{0,1,...,12\}$). In the following, each time point $s$ for which a prediction is calculated is called a landmark. Three methods are examined for predictions in this dynamic setting. As a first benchmark, we develop multiple logistic regression models with elastic net regularization. We then adopt a simple version of the landmarking approach for prediction which results in a single Cox proportional hazards model with time-dependent covariates14. As a third model, we build a Dynamic Bayesian Network (DBN) with the appealing property that all covariates observed up to time $s$ influence the current prediction.
2. Related Work

Much research has been dedicated to the prediction of mortality for dialysis patients. Table 1 summarizes prior work that differ by chosen method, set of covariates, landmark s and prediction horizon w. One popular approach is the application of logistic regression to data available at initialization of dialysis9, 10, 12. To enhance adoption in clinical practice, derived odds ratios are often converted to a simple additive risk score10, 12. A second approach applies Cox proportional hazards models but choose different landmarks for their analysis5, 6, 11. Geddes et al. use data at the start of dialysis and compare a Cox model, a C5.0 rule based model and a modified Charlson Comorbidity Index (CCI)9. In contrast, Cohen et al. use snapshot data of prevalent dialysis patients, not distinguishing by time on dialysis11. They note that the performance of their model decreases for longer prediction horizons because the patient condition may have changed over time. Floege et al. utilize information on incident patients captured in the first 90 days on dialysis and transform the regression coefficients to a simple risk score6. Although trained with incident patients, they apply their model to a large external data set including many prevalent dialysis patients. The performance expressed as AUC decreases from 0.76 to 0.73 for prevalent patients. A further stream of research concentrates on methods besides classical regression models. Kusiak et al. apply rough-set theory and decision trees to obtain 16 rule-based classifiers, which they combine using a voting scheme7. Knorr et al. compare a Support Vector Machine, a rule based learner and a boosted meta classifier8. Their focus on the temporal structure of dialysis data is similar to ours, but they concentrate on methods to extract temporal features from time series and integrate those features into their classifiers. The landmark is fixed at 100 days after start of dialysis. In both studies, the chosen prediction horizon is close to the median survival time of the analyzed population. Thus, the proportions of the two outcome classes are balanced, which is advantageous for the applied methods.

Table 1. Exemplary prior research on models to predict mortality of dialysis patients

<table>
<thead>
<tr>
<th>Authors</th>
<th>Features</th>
<th>Sample size</th>
<th>Method(s)</th>
<th>Landmark</th>
<th>Horizon</th>
<th>Performance (out-of-sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Couchoud et al. (2009)</td>
<td>9</td>
<td>4142</td>
<td>Logistic regression</td>
<td>Start of HD</td>
<td>6 months</td>
<td>AUC: 0.70</td>
</tr>
<tr>
<td>Mauri et al. (2008)</td>
<td>10</td>
<td>5738</td>
<td>Logistic regression</td>
<td>Start of HD</td>
<td>1 year</td>
<td>AUC: 0.78</td>
</tr>
<tr>
<td>Cohen et al. (2010)</td>
<td>5</td>
<td>876</td>
<td>Cox model</td>
<td>No diff.</td>
<td>6 months</td>
<td>AUC: 0.80</td>
</tr>
<tr>
<td>Floege et al. (2015)</td>
<td>15</td>
<td>9722</td>
<td>Cox model</td>
<td>0-90 days</td>
<td>1 year, 2 years</td>
<td>AUC: 0.73 (all pat., 1 yr.), 0.76 (incident pat., 1 yr.)</td>
</tr>
<tr>
<td>Kusiak et al. (2005)</td>
<td>&gt; 50</td>
<td>188</td>
<td>Rough set, Decision tree</td>
<td>After 15-20 sessions</td>
<td>3 years</td>
<td>Accuracy: 0.76 - 0.84</td>
</tr>
<tr>
<td>Knorr et al. (2009)</td>
<td>&gt; 40</td>
<td>&gt; 1000</td>
<td>SVM, Rule learner, adaBoost</td>
<td>First 100 sessions</td>
<td>5 years</td>
<td>Accuracy: SVM: 0.73, Rule learner: 0.70, adaBoost: 0.76</td>
</tr>
</tbody>
</table>

Independent of the application to dialysis care, several methods exist that allow dynamic predictions or integrate time-dependent covariates. The Cox model with time-dependent covariates is one popular option to integrate longitudinal observations into the analysis of survival data. However, this method can mainly be used for exploring the relationship between time-dependent covariates and the event of interest. The usage as a predictive model is limited unless the distribution of future covariates is known15, 16. As one solution, van Houwelingen proposes a landmarking approach to extend Cox regression models for dynamic predictions19 and suggests different modeling techniques to account for the dependency of the prediction on the landmark s. Aside from multistate models and joint models, Dynamic Bayesian Networks (DBNs) are capable of dynamic predictions and have been applied to several medical problems17-20. In the area of dialysis, a DBN was proposed for estimation of the patient’s dry weight21. DBNs provide a simple approach to integrate expert knowledge, handle missing values, and allow complex interactions of covariates by specifying joint probability distributions. However, to the best of our knowledge, neither of these dynamic methods has been utilized yet to predict mortality of dialysis patients.

3. Data

The data set covers seven dialysis clinics managed by a community nephrology practice in Western Pennsylvania. It was gathered between January 2006 and February 2014 and includes electronic health records of 1,484 hemodialysis patients who started their treatment at one of the seven clinics. Records include demographic data, clinical data such
as laboratory values, medications, treatments, diagnoses and hospitalizations, as well as data from each dialysis session. Since logistic regression models cannot handle censored observations, patients who were not followed until the end of the prediction horizon are excluded for this model. More than 90 features are extracted and 41 features are retained that are considered relevant by the nephrologists and have an acceptable coverage with less than 20% missing values. Corrupt values and outliers are removed after inspection of each variable’s distribution.

For each patient, data is captured at multiple time points: demographics and known diagnoses are usually entered at the onset of dialysis. Laboratory tests are drawn weekly, monthly or irregularly, depending on the type of laboratory test. At each dialysis session, additional data is recorded. In order to convert those values to a homogenous format and to cushion the influence of measurement errors, they are binned to aggregated values over 30-day periods. Binary features are used for all diagnoses and indicate any history of hypertension, hypotension, anemia, diabetes and a Charlson Comorbidity Index (CCI) > 2 before the end of the 30-day period. In contrast, pneumonia represents a short-term critical condition and the associated binary feature only captures a recent diagnosis within the period. Mean values are used for all laboratory tests, duration of dialysis session, interdialytic weight gain and Body Mass Index (BMI). Binary features for medications depict any prescription within the 30-day period. The occurrence of any complication, edema or extreme pre-dialysis systolic blood pressure (SBP) is represented similarly. For each period, the counts of hospitalizations, no shows and sessions with decreased duration due to ‘patient request’ are modeled with numeric features. The type of vascular access is determined by the one that is mostly used within the 30-day period. We use the first date of dialysis as a reference point for each patient and start with a period of 30 days around this date to capture all relevant information at initialization of dialysis. After initialization, consecutive periods of 30 days are employed until the end of observation. Note that the last observed period of a patient might be shorter than 30 days because the censoring event (either death or end of follow-up) occurs within that period. Since this could bias some of the aggregated measures (e.g. observing complications is less likely in shorter periods), observations in the last period were not used. This is assured by only predicting survival for patients still at risk at the end of a period and using the aggregated measures of the previous period(s).

Table 2. Characteristics of cohort with 1,484 patients in the first month on dialysis

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Mean (SD)</th>
<th>Occur.</th>
<th>Category</th>
<th>Feature</th>
<th>Occur.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demo-graphics</strong></td>
<td>Sex = Male</td>
<td></td>
<td>57%</td>
<td><strong>Miscellaneous</strong></td>
<td>BMI &lt; 20</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>69.1 (13.8)</td>
<td></td>
<td></td>
<td>Pre-ESRD care</td>
<td>78%</td>
</tr>
<tr>
<td><strong>Laboratory tests</strong></td>
<td>Hemoglobin (g/dL)</td>
<td>10.3 (1.2)</td>
<td></td>
<td><strong>Diagnoses</strong></td>
<td>Adj. CCI &gt; 2</td>
<td>16%</td>
</tr>
<tr>
<td></td>
<td>Kt/V</td>
<td>1.5 (0.4)</td>
<td></td>
<td></td>
<td>Hypertension</td>
<td>76%</td>
</tr>
<tr>
<td></td>
<td>Albumin (g/dL)</td>
<td>3.4 (0.5)</td>
<td></td>
<td></td>
<td>Hypotension</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Potassium (mmol/L)</td>
<td>4.3 (0.5)</td>
<td></td>
<td></td>
<td>Anemia</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>Creatinine (mg/dL)</td>
<td>4.7 (1.9)</td>
<td></td>
<td></td>
<td>Pneumonia</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Calcium (mg/dL)</td>
<td>8.6 (0.7)</td>
<td></td>
<td></td>
<td>Diabetes</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Transferrin sat. (%)</td>
<td>20.2 (10.8)</td>
<td></td>
<td><strong>Complication</strong></td>
<td>Poor access flow</td>
<td>6%</td>
</tr>
<tr>
<td></td>
<td>Phosphorus (mg/dL)</td>
<td>4.3 (1.2)</td>
<td></td>
<td></td>
<td>Access problem</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Dialysis session</strong></td>
<td>Edema</td>
<td></td>
<td>74%</td>
<td></td>
<td>Cramps</td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td>Pre-dialysis systolic blood pressure &gt; 180</td>
<td></td>
<td>2%</td>
<td></td>
<td>Shortness of breath</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Pre-dialysis systolic blood pressure &lt; 110</td>
<td></td>
<td>12%</td>
<td></td>
<td>Infiltration of needle</td>
<td>5%</td>
</tr>
<tr>
<td></td>
<td>Actual duration (hours)</td>
<td>3.8 (0.4)</td>
<td></td>
<td></td>
<td>Complaints about pain</td>
<td>6%</td>
</tr>
<tr>
<td></td>
<td>Interdialytic weight gain (kg)</td>
<td>1.5 (1.1)</td>
<td></td>
<td></td>
<td>Hypotension problem</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>#Duration decreased</td>
<td>0.6 (1.0)</td>
<td></td>
<td><strong>Medication</strong></td>
<td>Beta blockers</td>
<td>49%</td>
</tr>
<tr>
<td><strong>Miscellaneous</strong></td>
<td>#Hospitalizations</td>
<td>0.2 (0.5)</td>
<td></td>
<td></td>
<td>Anticoagulants</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td>#No shows</td>
<td>0.1 (0.3)</td>
<td></td>
<td></td>
<td>ESA</td>
<td>85%</td>
</tr>
<tr>
<td></td>
<td>Type of vasc. access = AVF or AVG</td>
<td></td>
<td>27%</td>
<td></td>
<td>Statins</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>Autonomy degree (normal, limited, special care)</td>
<td></td>
<td>44%, 37%, 18%</td>
<td></td>
<td>Antihypertensive medication</td>
<td>51%</td>
</tr>
</tbody>
</table>
Table 2 describes the population in the first month on dialysis with respect to the chosen features. In contrast to patient characteristics reported by the United States Renal Data System (USRDS), we have an older population (+ 6.1 years) with slightly higher mean values for albumin (+ 0.2 g/dL), hemoglobin (+ 0.4 g/dL) and a higher prevalence of fistulas and grafts (+ 8%)\(^4\). We excluded CKD and diabetes from the calculation of the Charlson Comorbidity Index (CCI) due to the high prevalence in our population. Further, prescriptions of ACE-inhibitors, Calcium channel blockers and Clonidine are combined in the feature ‘antihypertensive medication’. The feature ‘autonomy degree’ is derived from the patient’s mobility recorded at each arrival at/departure from a dialysis session and the living situation. Autonomy of patients in a wheelchair or geriatric chair is defined as ‘limited’. Patients in a bed, on a stretcher or living in a nursing home are labeled as ‘special care’. The feature depicts the worst autonomy degree observed in the 30-day period. We also include an indicator for the patient’s pre-ESRD care, which is defined as having any appointment at the community nephrology practice during earlier stages of CKD.

4. Methods

Regularized logistic regression

As a first approach, one logistic regression model is fitted for each landmark \(s\). Although the number of models generated by this approach might be problematic in clinical practice, it serves as a solid benchmark for comparison. The cohort is reduced to patients who are still at risk at \(s\) and for whom the individual outcome \(y_i\) at \(s + w\) is known. The aggregated covariates \(X_i\) of the previous 30 days are used at each landmark. Observations with missing values are excluded from the analysis. The resulting subset consists of 819 patients at \(s = 1\) and monotonically decreases to 449 patients at \(s = 12\). The one-year-mortality rate decreases from 27% at \(s = 0\) to 23% at \(s = 9\) and remains stable at that level.

We use elastic net regularization to obtain sparse models and potentially better prediction performance through the bias-variance trade-off\(^{22}\). The penalty term is a compromise between the \(L_1\) penalty of lasso and \(L_2\) penalty of ridge regression. Highly correlated variables tend to be selected or dropped as a group, which appears advantageous in our setting\(^{22}\). Since optimization over \(\alpha\) values did not show a significant improvement, we use equal weights for both penalties with \(\alpha = 0.5\). Coefficients are estimated at each landmark by maximizing the penalized log-likelihood

\[
LL_{pen} = n^{-1} \sum_{i=1}^{n} y_i \log(P(X_i)) + (1 - y_i) \log(1 - P(X_i)) - \lambda \sum_{j=1}^{p} \alpha \beta_j^2 + (1 - \alpha) |\beta_j|
\]

for \(n\) individuals and \(p\) covariates. The value of \(\lambda\) is chosen by iterating over a sequence of values and the one with the highest performance is selected\(^{23}\). As a performance measure, we repeat 5-fold cross validation ten times and obtain the average area under the ROC curve (AUC). The alternative approach of choosing the largest \(\lambda\) that results in an AUC within one standard error of the best AUC is tested for comparison (1-SE rule).

Cox proportional hazards model with landmarking

Cox models with time-dependent covariates provide the opportunity to explore the relationship between the event time \(T\) and time-dependent covariates \(X(t)\). However, the ability to predict is very limited after introducing time-dependent covariates unless the future distribution of \(X(t)\) is known\(^{15, 16}\). The landmarking approach described by van Houwelingen is one method to overcome this limitation by fitting standard proportional hazard models for different landmarks \(s\)\(^{12}\). Each of the simple Cox models is fitted to the set of patients still at risk at \(s\), using the associated covariate values \(X(s)\) and censoring all events after \(s + w\). The hazard function is given by

\[
h(t|X(s), s, w) = h_0(t|s, w) \exp(X(s)\beta(s)) \quad \text{for} \quad s \leq t \leq s + w.
\]

Each corresponding data set is called a ‘landmark data set’. Several methods are proposed to consolidate the multiple landmark models with different coefficients \(\beta(s)\) to one single ‘super model’ with smooth landmark-dependent coefficients \(\beta(s)\)\(^{15}\). In this work, we use a simple ‘super model’ by combining all landmark data sets to one stacked data set and stratifying the Cox model by \(s\) to obtain landmark-specific baseline hazards. As a consequence, the coefficients are the same for all landmarks, but the estimates are based on landmark-dependent baseline hazards. For simplicity, we do not include further interaction terms of the covariates with the landmark, which would yield truly landmark-dependent coefficients. The sandwich estimators of Lin and Wei are used to derive robust standard errors for the estimated coefficients\(^{24}\). We estimate the model on the dialysis data and include right-censored patients. The stacked landmark data set consists of 1,219 patients and 9,763 complete observations for different landmarks. Furthermore, only features selected in at least one regularized logistic regression model are used for this analysis. Ten times 5-fold cross validation is conducted to obtain robust estimates of the AUC.
Dynamic Bayesian Network

A Bayesian Network (BN) is defined by a pair \((G, P)\) where \(G\) is a directed acyclic graph and \(P\) specifies the joint probability distribution of a set of random variables \(X = \{X_1, ..., X_n\}\). The nodes of \(G\) correspond to the random variables and the directed edges \(E \subseteq N \times N\) express dependencies between the variables. A BN provides a compact representation of the joint distribution by factorization into conditional probability distributions (CPDs)

\[
P(X) = \prod_{j \in N} P(X_j | \pi(X_j))
\]

where \(\pi(X_j)\) denotes the parents of \(X_j\) in \(G\). Random variables are allowed to be hidden and the associated CPDs can be of arbitrary form. If realizations of some random variables are observed, probabilistic inference can be used to update the beliefs about the state of any other \(X_j\).

A Dynamic Bayesian Network is a temporal extension of a BN and allows the modeling of probability distributions over semi-infinite collections of random variables \((Z_0, Z_1, ...)\). Thus, a set of random variables can be represented conveniently at different time points and stochastic processes can be modeled. Time is discretized into distinct time slices \(t\) and for each time slice a set of random variables is given by \(Z_t\). Dependencies can be modeled within each time slice (local arcs) and between time slices (temporal arcs). Usually, the Markov property is assumed and the parents of a node \(j\) in \(t\) (i.e. \(\pi(Z_{j,t})\)) have to be located in the same or previous time slice. If \(Z_{t-1}\) is a parent of \(Z_t\), this node is called persistent. Furthermore, the structure of the network is assumed to be invariant for all time slices. A DBN is then defined by a pair \((B_0, B_\infty)\) where \(B_0\) is a BN and \(B_\infty\) is a two-slice temporal Bayes net. \(B_0\) defines the initial distribution of \(Z_0\) at \(t = 0\) and \(B_\infty\) specifies the distribution of the random variables for subsequent time points by

\[
P(Z_{t+1}|Z_t) = \prod_{i \in N} P(Z_{t+1,i} | \pi(Z_{t+1,i}))
\]

where \(Z_{t+1}\) is the \(i\)-th node in time slice \(t\). It is always possible to “unroll” a DBN to a BN for a fixed number of time slices and apply the common inference algorithms. However, computational complexity can be decreased if algorithms are applied that exploit the repetitive structure of a DBN. Both structure and parameters of the CPDs can either be learned from data or specified using expert knowledge.

In this study, we model a DBN with discrete nodes and time slices for each 30-day period. The initial network structure is manually defined after extensive discussions with nephrologists, and parameters are learned from data. One big challenge is the number of parameters that need to be estimated from a limited number of observations. For each node, the size of the conditional probability table grows exponentially with the number of parents and number of parents’ states. Thus, we reduce the number of features to those that are considered highly relevant by the experts and have shown to have significant influence on mortality in prior research. Since our goal is prediction of mortality, we include one observable node ‘death’ and a hidden, persistent node ‘mortality risk’. The former depicts the separation of risk factors by such hidden variables reduces the number of parameters significantly. Additionally, some simplified but useful insights can be derived from the distribution of these abstract concepts when we want to explore the reasoning behind a survival prediction. More than thirty further network structures are evaluated with slight variations in modeled nodes and arcs. For example, we add persistencies for some risk factors that are relatively stable over time. Modeling such persistencies allows a more detailed forecast of the patient’s risk factors in future periods. For example, if we observe a patient with a low BMI at landmark \(s\) and the learned transition probability \(P(\text{Low BMI}_{t+1} = \text{true}|\text{Low BMI}_t = \text{true})\) is high, the belief about observing a low BMI at \(s+1\) will be high and will impact the mortality risk for that period. Parameters

\[
P(T_{\text{surv},i} > s + w | T_{\text{surv},i} > s) = \prod_{t=s}^{s+w-1} P(Z_{\text{death},t} = \text{false})
\]

Two further hidden and persistent variables are included that represent two major risks of mortality: ‘relative cardiovascular risk’ and ‘relative infection risk’. This approach follows that of Cornalba et al. who model similar relative risks for dialysis patients in a static BN. The separation of risk factors by such hidden variables reduces the number of parameters significantly. Additionally, some simplified but useful insights can be derived from the distribution of these abstract concepts when we want to explore the reasoning behind a survival prediction. More than thirty further network structures are evaluated with slight variations in modeled nodes and arcs. For example, we add persistencies for some risk factors that are relatively stable over time. Modeling such persistencies allows a more detailed forecast of the patient’s risk factors in future periods. For example, if we observe a patient with a low BMI at landmark \(s\) and the learned transition probability \(P(\text{Low BMI}_{t+1} = \text{true}|\text{Low BMI}_t = \text{true})\) is high, the belief about observing a low BMI at \(s+1\) will be high and will impact the mortality risk for that period. Parameters
are learned offline from data by maximizing the log-likelihood using the expectation–maximization (EM) algorithm since some of the variables are hidden\(^3\). We initialize the CPDs with manually chosen parameters reflecting our beliefs and results of prior work. We train the DBN on the dialysis data set and exclude all observations after \(s = 36\) to strengthen the focus on risk factors at the beginning of dialysis. In contrast to the previous two methods, missing values can be handled natively by the DBN and we utilize all 23,292 observed patient-periods from 1,484 patients for parameter learning. Nevertheless, cases with an unknown outcome, i.e. right-censoring between \(s\) and \(s + w\), have to be excluded from the evaluation of the DBN. Cases with missing covariate values are also excluded from the evaluation to enhance comparability with the other two models.

### 4. Results

Regularization of the logistic regression models leads to thirteen sparse models. On average, 18 of the 42 features are selected but the feature set differs considerably between landmarks. We highlight four observed patterns: features included consistently for most landmarks, never selected, only included for early landmarks and unstable selection. Seven risk factors are consistently found for at least ten landmarks: higher age, adj. CCI > 2, lower albumin, no antihypertensive medication, low pre-dialysis systolic blood pressure (SBP), lower autonomy degree and catheter as vascular access are associated with a lower probability of one-year survival. Most of the results are in line with existing research and expectations of nephrologists, though it might be counterintuitive that a low pre-dialysis SBP is a strong predictor of mortality. The covariate is included in 11 out of 13 periods with an average odds ratio of 1.55. In contrast, high pre-dialysis SBP shows a similar average odds ratio (1.44) but is only included in three periods. We test different thresholds but gain similar results. The effect is also observed in other studies and this controversy is discussed under the term “reverse epidemiology”\(^31, 32\). One hypothesis is that hypotension is a sign of an already severe state of a cardiovascular co-morbidity (e.g. congestive heart failure). Six features are not selected for any landmark: diagnosis of hypotension, phosphorus, transferrin saturation, calcium * phosphorus, number of no shows and number of sessions with decreased duration by patient request.

Many covariates are only included for few landmarks and interpretation is more difficult. The features ‘beta blockers’, ‘anticoagulants’, ‘BMI < 20’, ‘calcium’, ‘edema’ and ‘pre-ESRD care’ are mostly selected at early landmarks and might indicate a systematic change of predictors over time. For ‘pre-ESRD care’, the diminishing influence may be explained by the circumstance that pre-ESRD care has a big impact on patient’s condition at the start of dialysis (e.g. installed fistula, stable biomarkers, patient education), but these differences with other patients are likely to diminish over time because i) the physicians make up for missed preparation quickly (e.g. installation of a fistula), and ii) particularly fragile patients without pre-ESRD care may die within the first few months.

For a third group of features, a highly unstable selection pattern is observed. For example, ‘pneumonia’ is included in the models for \(s \in \{1, 4, 5, 6, 11\}\), though with a relatively high average odds ratio of 2.97. In this case, the limited number of observations is a potential explanation: with less than ten instances in most periods, the robustness of estimation may be impaired. However, for most other features with an unstable pattern, this explanation does not hold. Table 3 depicts representative features for each pattern and associated coefficients for seven landmarks. Even though inspecting coefficients provides useful insights, it is important to note that regularized estimates are biased in favor of lower variance. The models show an acceptable discriminative ability with an average AUC over all landmarks of 0.750. The performance decreases for models at later landmarks (Figure 2). Choosing \(\lambda\) according to the 1-SE rule results in sparser models (on average 13 features) and an average AUC of 0.745.

**Table 3.** Estimated non-zero coefficients of the logistic regression models for exemplary features

<table>
<thead>
<tr>
<th>Observed selection pattern</th>
<th>Exemplary features</th>
<th>Landmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Consistent</td>
<td>Albumin</td>
<td>-0.57</td>
</tr>
<tr>
<td></td>
<td>Pre-dialysis SBP &lt; 110</td>
<td>0.33</td>
</tr>
<tr>
<td>Early landmarks</td>
<td>BMI &lt; 20</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Pre-ESRD care</td>
<td>-0.56</td>
</tr>
<tr>
<td>Unstable</td>
<td>Pre-dialysis SBP &gt; 180</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pneumonia</td>
<td></td>
</tr>
<tr>
<td>Number of patients</td>
<td>713</td>
<td>674</td>
</tr>
</tbody>
</table>

1789
Many effects observed in the logistic regression models are reinforced by the Cox model with landmarking: age, sex, adj. CCI > 2, albumin, potassium, low pre-dialysis SBP and autonomy degree are significantly associated with mortality at a level of $\alpha = 0.05$. Additionally, the Cox model identifies beta blockers, shortness of breath, complaints about pain, edema, hospitalizations and pre-ESRD care as further significant risk factors. The estimated coefficients of all significant covariates are reported in Table 4. The discriminative performance ranges between 0.809 at $s = 2$ and 0.721 at $s = 11$ and is slightly better compared to logistic regression at each landmark (Figure 2). However, a similar trend of decreasing performance for later landmarks is observed.

Table 4. Significant coefficients of the Cox model with landmarking

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>$\beta$</th>
<th>$\exp(\beta)$</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Age</td>
<td>0.039</td>
<td>1.039</td>
<td>0.006</td>
<td>$&lt; 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Sex = male</td>
<td>0.361</td>
<td>1.435</td>
<td>0.125</td>
<td>0.004</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>Adj. CCI &gt; 2</td>
<td>0.357</td>
<td>1.429</td>
<td>0.131</td>
<td>0.006</td>
</tr>
<tr>
<td>Laboratory tests</td>
<td>Albumin</td>
<td>-0.847</td>
<td>0.428</td>
<td>0.118</td>
<td>$&lt; 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Potassium</td>
<td>0.289</td>
<td>1.335</td>
<td>0.075</td>
<td>$&lt; 10^{-4}$</td>
</tr>
<tr>
<td>Medication</td>
<td>Beta blockers</td>
<td>-0.270</td>
<td>0.763</td>
<td>0.113</td>
<td>0.017</td>
</tr>
<tr>
<td>Complications</td>
<td>Shortness of breath</td>
<td>0.463</td>
<td>1.589</td>
<td>0.195</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Complaints about pain</td>
<td>0.243</td>
<td>1.275</td>
<td>0.123</td>
<td>0.049</td>
</tr>
<tr>
<td>Dialysis session</td>
<td>Edema</td>
<td>0.215</td>
<td>1.240</td>
<td>0.092</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>Actual duration</td>
<td>-0.314</td>
<td>0.730</td>
<td>0.144</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Pre-dialysis SBP &lt; 110</td>
<td>0.406</td>
<td>1.501</td>
<td>0.118</td>
<td>0.001</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Autonomy degree: limited</td>
<td>0.390</td>
<td>1.476</td>
<td>0.112</td>
<td>$&lt; 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Autonomy degree: special care</td>
<td>0.656</td>
<td>1.927</td>
<td>0.147</td>
<td>$&lt; 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Pre-ESRD care</td>
<td>-0.461</td>
<td>0.630</td>
<td>0.145</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>#Hospitalizations</td>
<td>0.237</td>
<td>1.267</td>
<td>0.051</td>
<td>$&lt; 10^{-4}$</td>
</tr>
</tbody>
</table>

After testing more than thirty different network structures, the DBN depicted in Figure 1 demonstrates the best average discrimination for the prediction horizon of one year (AUC = 0.745). Whereas short-term predictions yield a good performance (AUC = 0.86 for $w = 1$), a longer horizon of $w = 12$ leads to decreased AUCs (0.704 - 0.782). Hence, the DBN is inferior to the other two models for most landmarks (Figure 2). Adding more risk factors does not result in a better AUC since it increases the number of model parameters substantially and impairs the robustness of learned parameters.

Figure 1. One time-slice of a DBN with hidden nodes in grey and inter-slice dependencies depicted as dashed arcs.

Figure 2. Average AUC and standard error for different landmarks and a horizon of $w = 12$. 
A qualitative validation of the network was conducted by discussing the results for exemplary patients with nephrologists and how they would use it in the clinics. Two simplified cases are depicted in Figures 3 and 4. Both patients undergo dialysis for four periods ($s = 4$) and several adverse observations are noted in period four: Hemoglobin < 10, Potassium $\geq$ 6, Low SBP = true and Edema = true. Whereas Patient A has already exhibited the same risk factors in the previous periods, Patient B has been in good condition until period four (Hemoglobin 10-11, Potassium < 6, Low SBP = false, Edema = false). As depicted in Figure 3, the inferred probabilities for the node ‘relative risk (cardiovascular)’ are on a very different level for both patients in periods one to three and converge in period four. However, probabilities in period four are not identical (0.82 vs. 0.86), because the previous periods still have a small influence on current beliefs. The difference is also reflected in the probability of death (Figure 4), but fades away for future periods where patients’ states are more uncertain. This simple example illustrates two important characteristics of the DBN: past observations still impact beliefs at later landmarks and the random variables converge to a steady state for periods in the far future. In ongoing research, we are identifying and evaluating other clinically relevant and useful use cases based on actual patient data.

5. Discussion

In this study, we explored three different approaches to dynamically predict mortality of dialysis patients. Using observational data for model building is challenging and one should be particularly vigilant to avoid biases. On the current data set, the adjusted Cox model achieved the best discriminative performance. Nevertheless, further model properties should be considered to gain a comprehensive view of the opportunities and limitations of each approach. Regularized logistic regression might be advantageous if there are many (potentially correlated) features and the number of landmarks is small. However, the number of required models increases with landmarks and prediction horizons that are of interest. Usability in a clinical setting might therefore be impaired. One further challenge arises from the separate estimation of the models: selected features might be substantially different for two consecutive landmarks and hinder interpretability. A joint estimation of the models that encourages similar sparsity patterns in all models is an interesting extension that would mitigate this effect$^{33}$. Using a Cox model with landmarking can be a compelling alternative if many observations are censored and one single model needs to be derived. The dependency of the prediction on the landmarks can be achieved by employing simple modifications of the data set and the standard Cox model. In contrast to this work, more complex dependencies between landmark and covariate can also be modeled. Instead of binary survival prediction for a fixed horizon, the estimated models can produce informative survival curves over a complete period of interest. Yet, only the most recent observations are considered at each landmark and modeling many landmark-dependent coefficients leads to a large number of parameters. In contrast, a DBN incorporates all previous observations, allows complex interactions of variables, handles missing values and expert knowledge can be integrated easily. Once the DBN is trained, the prediction horizon can be chosen flexibly and it is possible to perform inference for any random variable included in the network (e.g. hospitalizations). The rapidly increasing number of parameters is one major challenge that limits the number of modeled risk factors and the number of bins for discretizing continuous variables. It is difficult to find a reasonable trade-off between the detailed representation of risk factors and parameters that can be learned from small data sets. Modeling techniques that reduce the number of parameters (e.g. Noisy-OR gates) may provide an interesting extension to our approach but introduce further assumptions on causal dependencies between risk factors$^{34}$. 

![Figure 3](image3.png) Relative risk (cardiovascular) for two patients with identical observations in period four, but different observations in previous periods.

![Figure 4](image4.png) Probability of death for Patient A and B in future periods.
The appropriate representation of long-term risks is also challenging and affects the performance of the DBN in this work. A continuous time BN could be an interesting alternative to our approach, since it would allow for completely flexible prediction horizons. In future work, we plan to test the performance of our models on a completely separate data set to assess generalization capabilities.

Three clinically relevant use cases have been identified in the discussions with nephrologists. First, at the onset of dialysis, patients are often anxious to know the prognosis of their condition. An individualized prediction has the potential to facilitate shared decision making and help in the communication with the patient. Second, once patients are on dialysis, their compliance is often insufficient. Individualized predictions with different scenarios can help to improve compliance with interventions, such as patient education, and to present potential consequences of their own decisions (e.g. a catheter leading to higher mortality risk). Third, the model can be used by clinicians for risk stratification of their patient population and identification of patients with elevated risk profiles. Future work will investigate the implementation and evaluation of the prediction models with their significant risk factors in the actual care delivery setting.

6. Conclusion

Patients on hemodialysis are in a vulnerable health condition and robust risk prediction models can support clinicians to identify and target high risk patients. The focus of this work is on predicting mortality but the methods can easily be adapted for other events of interest (e.g. hospitalizations or infections). In contrast to prior work, we explore three approaches to incorporate the temporal nature of EHR to dynamically adjust predictions when patients progress in time. The explicit representation of the time dimension is methodologically challenging, but offers great potential to exploit available EHR data and to represent changing patient characteristics more appropriately. We use variations of two common statistical methods, logistic regression and Cox regression, to allow dynamic predictions at different landmarks. Furthermore, using a machine learning approach, we build a DBN that allows complex interactions between risk factors and utilizes the complete patient history to predict future risks. The Cox model achieves the best discrimination performance with an average AUC of 0.767. The regularized logistic regression models are slightly inferior and use separate models for each landmark (average AUC: 0.750). The DBN demonstrates a good performance for short prediction horizons, but cannot achieve similar results for the horizon of one year (average AUC: 0.745). The large volume of data required for learning robust parameters of complex DBNs is one major challenge. The models provide interesting opportunities to monitor patients over time and can be used as a starting point for further extensions of dynamic models in the domain of dialysis care.

Acknowledgements

We are grateful to the physicians and staff of the nephrology practice that provided the data for this study, and Dr. A. Chouldechova at the Heinz College at CMU for many helpful suggestions and feedback on the study.

References


Multi-modal Patient Cohort Identification from EEG Report and Signal Data

Travis R. Goodwin, MS, Sanda M. Harabagiu, PhD
University of Texas at Dallas, Richardson, TX

Abstract

Clinical electroencephalography (EEG) is the most important investigation in the diagnosis and management of epilepsies. An EEG records the electrical activity along the scalp and measures spontaneous electrical activity of the brain. Because the EEG signal is complex, its interpretation is known to produce moderate inter-observer agreement among neurologists. This problem can be addressed by providing clinical experts with the ability to automatically retrieve similar EEG signals and EEG reports through a patient cohort retrieval system operating on a vast archive of EEG data. In this paper, we present a multi-modal EEG patient cohort retrieval system called MERCuRY which leverages the heterogeneous nature of EEG data by processing both the clinical narratives from EEG reports as well as the raw electrode potentials derived from the recorded EEG signal data. At the core of MERCuRY is a novel multi-modal clinical indexing scheme which relies on EEG data representations obtained through deep learning. The index is used by two clinical relevance models that we have generated for identifying patient cohorts satisfying the inclusion and exclusion criteria expressed in natural language queries. Evaluations of the MERCuRY system measured the relevance of the patient cohorts, obtaining MAP scores of 69.87% and a NDCG of 83.21%.

Introduction

Clinical electroencephalography (EEG) is an electrophysiological monitoring method used to record electrical activity of the brain. Clinical EEG is the most important investigation in the diagnosis and management of epilepsies and can also be used to evaluate other types of brain disorders. An EEG records the electrical activity along the scalp and measures spontaneous electrical activity of the brain. Unfortunately, as noted in by Beniczky et al., the EEG signal is complex and inter-observer agreement for EEG interpretation is known to be moderate. Such interpretations of EEG recordings are available in EEG reports. As more clinical EEG data becomes available, the interpretation of EEG signals can be improved by providing neurologists with results of search for patients that exhibit similar EEG characteristics. Searching the EEG signals and reports results in the identification of patient cohorts that inform the clinical decision of neurologists and enable comparative clinical effectiveness research. For example, a neurologist suspecting that her patient has epilepsy potential formulated the query (Q1) History of seizures and EEG with TIRDA without sharps, spikes, or electrographic seizures. When inspecting the EEG signals and reports of the resulting patient cohort, the neurologist was able to observe the specific features of the EEG for patients that exhibited epileptic potential. In another search instance, a neurologist researcher was interested in one of the research priorities for improving surveillance and prevention of epilepsy as reported by England et al., namely, to identify effective interventions for epilepsy accompanied by mental health comorbidities. This researcher formulated the query (Q2) History of Alzheimer and abnormal EEG. The patient cohort that was identified enabled the researcher to observe the treatment outcomes as well as the clinical correlations documented in the EEG reports.

To ensure that patients from a cohort satisfy the criteria expressed in the natural language queries formulated by neurologists, it is important to not only consider the narrative from EEG reports, but also the EEG signal data. Searching for patient cohorts by considering EEG signals and reports relies on (1) an index of the EEG clinical information and (2) relevance models that identify records of relevant patients against a query. Indexing EEG clinical information requires organizing both narratives from the EEG reports and signal data from the EEG recordings. Consequently, the EEG index needs to capture multi-modal clinical knowledge processed both from the reports and the signal recordings. While medical language processing enables the indexing of information form the EEG reports, the index must also comprise a representation of EEG signal recordings. In addition, the relevance models used by the patient cohort retrieval system must account for inclusion and exclusion criteria inferred from processing the natural language query. To address these problems, we have developed a patient cohort retrieval system which produces a multi-modal index of big EEG data. We have also implemented two relevance models to identify the most relevant patients based on their EEG reports and also based on the properties of the EEG signal recordings. The patient cohort retrieval system, called MERCuRY (Multi-modal EncephalogRam patient Cohort discoveryRY), uses medical language processing to identify the inclusion and exclusion criteria from the queries and to index the clinical knowledge from the EEG reports. In addition, MERCuRY has two novel aspects not present in previous approaches for patient cohort retrieval: (1) it uses deep learning to represent the EEG signal and to produce a multi-modal EEG index; and (2) it operates based on two EEG relevance models – one that uses only the clinical information from the EEG reports, and
a second one which also considers the EEG signal information. We evaluated MERCuRY by using expert judgements of queries against a collection of nearly 20,000 EEGs.

**Background**

The ability to automatically identify patient cohorts satisfying a wide range of criteria – including clinical, demographic, and social information – has applications in numerous use cases, as pointed out in by Shivade et al. including (a) clinical trial recruitment; (b) outcome prediction; and (c) survival analysis. Although the identification of patient cohorts is a complex task, many systems aiming to resolve it automatically have used statistical techniques or machine learning methods taking advantage of natural language processing (NLP) of the clinical documents. However, these systems cannot rank the identified patients based on the relevance of the patient to the cohort criteria. This notion of relevance is at the core of information retrieval (IR) systems. Thus, viewing the problem of patient cohort identification as an IR problem enables us to not only identify which patients belong to a cohort, but to also rank patients based on relevance to the inclusion and exclusion criteria used in the query.

Using information retrieval for patient cohort identification was considered in 2011 and 2012 by the Medical Records track in the annual Text REtrieval Conference (TREC) hosted by the National Institute for Standards and Technology (NIST). When patient cohort identification systems are presented with a query expressing the inclusion/exclusion criteria for a desired patient cohort, a ranked list of patients representing the cohort is produced where each patient may be associated with multiple medical records. Thus, identifying a ranked list of patients is equivalent to producing a ranked list of sets of medical records, each pertaining to a patient belonging to the cohort. In MERCuRY, we have adopted the same framework of identifying patients that are relevant to a cohort and ranking them according to their relevance to the given cohort criteria. However, unlike the TREC patient cohort retrieval systems, which considered only the clinical texts available from a large set of electronic medical records, MERCuRY uses a multi-modal index that encodes textual data available from EEG reports as well as signal data produced by EEG signal recordings.

Historically, the majority of multi-modal retrieval systems have operated on text and image data. For example, Demner-Fushman et al. designed a biomedical article retrieval system which allows users to not only discover biomedical articles relevant to a query, but to also discover similar images to those found in any retrieved articles. Their approach clusters images using a large number of visual features capturing color, edge, texture, and other image information. By contrast, our approach relies on unsupervised deep learning to generate a fingerprint of EEG data. As such, although designed for EEG data, our architecture can be easily adapted to support other types of (physiological) waveform data (such as ECGs). Other forms of physiological waveform data were previously investigated by the AALIM system which enabled cardiac decision support by allowing physicians to locate similar patients according to the ECG, echo, and audio data associated with the patient. However, unlike our approach, their system does not support search: it can only identify similar patients to provided ECG, echo, or audio recording and thus cannot be used to discover patients matching arbitrary criteria.

The MERCuRY system presented in this paper relies on medical language processing techniques that were informed by the experience gained from participating in the 2010, 2011 and 2012 Informatics for Integrating Biology and the Bedside (i2b2) Challenges on NLP for Clinical Records, which focused on the automatic identification of medical concepts and events in clinical texts. In contrast, the EpiDEA patient cohort identification system which also retrieves EEG-specific patient cohorts, operates on discharge summaries to recognize patients that are relevant only to some pre-defined queries obtained by relying on the EpSO ontology.

**Data**

The MERCuRY system was developed to identify patient cohorts from the big EEG data available from the Temple University Hospital (TUH) EEG Corpus (over 25,000 sessions and 15,000 patients collected over 12 years). This dataset is unique because, in addition to the raw signal information, physician’s EEG reports are provided for each EEG. Following the American Clinical Neurophysiology Society Guidelines for writing EEG reports, the EEG reports from the TUH corpus start with a CLINICAL HISTORY of the patient, describing the patient’s age, gender, and relevant medical conditions at the time of the recording (e.g., “after cardiac arrest”) followed by a list of the medications which may influence the EEG. The INTRODUCTION section is the depiction of the techniques used for the EEG (e.g. “digital video EEG”, “using standard 10-20 system of electrode placement with 1 channel of EKG”), as well as the patient’s conditions prevalent at the time of the recording (e.g., fasting, sleep deprivation) and level of consciousness (e.g. “comatose”). The DESCRIPTION section is the mandatory part of the EEG report, and it provides a description of any notable epileptiform activity (e.g. “sharp wave”), patterns (e.g. “burst suppression pattern”) and events (“very quick jerks of the head”). In the IMPRESSION section, the physician states whether the EEG readings are
normal or abnormal. If abnormal, then the contributing epileptiform phenomena are listed. The final section of the EEG report, the CLINICAL CORRELATIONS section explains what the EEG findings mean in terms of clinical interpretation\textsuperscript{15} (e.g. “very worrisome prognostic features”). Each EEG report in the TUH corpus is associated with the EEG signal recording it interprets. The signal information consists of 24 to 36 channels of signal data as well as an additional annotation channel providing markers identifying events of interest to physicians and technicians. EEG signals are sampled at a rate of 250 Hz or 256 Hz using 16-bits per sample. Each EEG recording from the TUH EEG corpus contains roughly 20 megabytes of raw data, stored in the European Data Format (EDF+) file schema\textsuperscript{16}.

Methods
MERCuRY is a multi-modal patient cohort discovery system which allows neurologists to inspect the EEG records as well as the EEG signal recordings of patients deemed relevant to a query expressing inclusion and exclusion criteria through natural language. As illustrated in Figure 1, the neurologist query is analyzed to identify the inclusion and exclusion criteria. The results of query analysis inform two different relevance models (illustrated as Case 1 and Case 2 in Figure 1) which rely on the multi-modal EEG index encoding information identified in the EEG reports and signal recordings. When EEG reports are indexed, the sections of the EEG reports are identified and medical language processing is performed to identify the terms and phrases of the dictionary and to create tiered inverted lists. When the EEG signal recordings are processed, they are represented by EEG signal fingerprints which are produced by deep learning methods. EEG signal recordings are converted into low-dimensional fingerprint vectors which are included in the multi-modal index. Additional details of the index are provided later in the paper. As shown in Figure 1, in MERCuRY we considered two relevance models designed to identify and rank patients based on their relevance to the patient cohort query: Case 1, in which the EEG signal fingerprints are ignored and only the EEG reports are used, and Case 2 in which both the EEG fingerprints and reports are used. These two cases allowed us to experimentally evaluate the impact of the EEG signal fingerprint representation on the overall performance of MERCuRY in identifying patient cohorts.

Figure 1: Overview of the MERCuRY Patient Cohort Discovery System

A. Indexing the EEG Big Data
The multi-modal index used by the MERCURY system organizes the information from the EEG reports as well as the information from the EEG signal recordings. It contains both a term dictionary and a medical concept dictionary, listing all the terms and medical concepts discerned from the EEG reports. We considered five medical concept types: (1) medical problems, (2) medical tests; (3) medical treatments (including medications); (4) EEG patterns and activities; as well as (5) EEG events. Because medical concepts often are multi-term expressions (e.g. “spike and slow waves”), the medical concept dictionary used term IDs to associate a concept with all terms expressing it (e.g. “spike and slow waves” is associated with the terms “spike”, “slow” and “wave”). Moreover, as illustrated in Figure 2, each entry from the term dictionary is linked to a pair of inverted lists: the first corresponding to positive polarity associated with the term while the second corresponding to negative polarity. By using polarity information (which is automatically processed from the medical language used in the EEG reports), we have designed a multi-tiered index. Each of the tiered inverted lists is implemented as a linked list. Each cell of those lists indicates for every occurrence of the term: (1) in which EEG report the term was observed; (2) the EEG signal fingerprint of that EEG report; (3) in which section of the EEG report was the term observed; (4) in which position of the EEG section; (5) whether the term belongs to a medical concept identified in the EEG report; and (6) if so, what position does the term have in the concept.
Figure 2: The Multi-Modal Tiered Index used in the MERCuRY Patient Cohort Discovery System

The EEG signal fingerprints are representations of the EEG signal recordings obtained through deep-learning techniques (described later in the paper) and organized in a similarity-based hierarchy which enables the discovery of relevant patients when the EEG signal recordings are also considered (Case 2 illustrated in Figure 1). When the EEG signal recordings are not used for identifying patient cohorts (Case 1 illustrated in Figure 1), only the term dictionary, the medical concept dictionary, and the tiered inverted lists from the index are used. Creating the multi-modal tiered index for the MERCuRY Patient Cohort Discovery System involves (1) the recognition of sections of the EEG reports; (2) medical language processing to determine (a) the terms from the dictionary, (b) their polarity and (c) the medical concepts; (3) generating the fingerprints for the EEG recording; and (4) organizing the EEG signal fingerprints in the similarity-based hierarchy.

Section Identification. Sections were identified through a rule-based section segmentation approach. Our rules were defined after manually reviewing 300 randomly sampled EEG reports. We detected a set of candidate headers by discovering all sequences of all capitalized words ending in a colon or line break, and normalized section titles based on simple regular expressions. For example, “description of the record”, “description of record”, and “description of the recording” would all be normalized to DESCRIPTION.

Medical Language Processing. In order to build (1) the term dictionary; (2) the medical concept dictionary and (3) the two polarity-informed posting lists we have used the following sequence of steps:

(Step 1) Tokenizing the EEG reports: we relied on Stanford’s CoreNLP pipeline to detect sentences and tokens from every EEG report.

(Step 2) Discovering the Dictionary Terms: Each token was normalized in order to account for any lexical variation (e.g. “waves” and “wave” or “markedly and “marked”) using Stanford’s CoreNLP lemmatizer. The resultant lemmatized terms formed the basis of the dictionary.

(Step 3) Identifying the Polarity: Term polarity was cast as a classification problem implemented as a conditional random field (CRF). Leveraging our previous experience with the i2b2 challenge, the CRF assigned a binary polarity value (i.e. positive or negative) to each term based on feature vector containing lexical information as well as information from external resources, such as the NegEx negation detection system, the Harvard General Inquirer, the Unified Medical Language System (UMLS) meta-thesaurus, and MetaMap. Specifically, we considered nine features: (1) the section name, (2) whether the term was considered a modifier by NegEx, (3) whether the term was within a NegEx negation span, (4) whether the term was in the ‘IF’ category of the Harvard General Inquirer, (5) the part-of-speech tag assigned to the token by Stanford’s CoreNLP part-of-speech tagger, (6) whether the term belonged to a UMLS concept, (7) whether the term belongs to a MetaMap concept, (8) the original cased term before lemmatization, and (9) the lowercased and lemmatized form of the term. The classifier was trained using 2,349 manual annotations.

(Step 4) Identifying Medical Concepts: An automatic system for medical concept recognition previously developed for the 2010 i2b2 challenge recognized medical problems, tests, treatments. In addition, we have produced 4,254 new annotations for EEG patterns and events as well as EEG activities and re-trained the concept recognizer to identify all these types of concepts. Concept extraction was cast as a classification task, in which a CRF was used to detect medical concept boundaries. A support vector machine (SVM) was used to classify each concept into one of five types: medical problem, medical test, medical treatment, EEG activity, or EEG event.
Generating Fingerprints of EEG Signal Recordings. In the TUH EEG corpus, the EEG signals are encoded as dense floating-point matrices of the form \( \mathbf{D} \in \mathbb{R}^{N \times L} \), where \( N \in [24, 36] \) is the number of electrode potential channels in the EEG and \( L \) is the number of samples (such that duration of the EEG recording in seconds is equal to \( L / 250 \) Hz). Thus \( \mathbf{D}_{ij} \) encodes the magnitude of the potential recording on the \( i \)-th channel during the \( j \)-th time sample. Both the number of channels and the number of samples vary not only across patients, but also across EEG recording sessions. These variations, particularly when considered with the large amount of data encoded in each matrix (typically 20 megabytes), make it difficult to not only characterize the relevance EEG signals to a particular patient cohort, but also to determine the similarity between two EEG signals. For example, consider that a single naïve pass over the TUH EEG corpus requires considering over 400 gigabytes worth of information. Consequently, we devised a representation of the EEG recordings that not only requires less memory, but enables rapid similarity detection between EEG signal recordings. This allows us to compactly encode the information from the EEG signals (reducing 20 megabytes of signal data to a few hundred bytes). Our representation is based on EEG signal recording fingerprints obtained with a recurrent neural network. Recursive neural networks are deep learning architectures which enabled us to generate fingerprints for each EEG in the TUH EEG corpus in a matter of hours instead of weeks. Because traditional neural networks have difficulty operating on sequential data (e.g. EEG signals) as they cannot consider relationships between successive inputs (e.g. between successive samples), we have used a recurrent neural network\(^{15} \) (a neural network with a loop) which allowed information learned at each sample to persist between samples in the same EEG signal. The learned information (known as the internal memory state) is updated with each sample until ultimately becoming the fingerprint for that EEG recording. Figure 3 illustrates the recurrent neural network used for EEG fingerprinting. As shown, the unrolled network (on the right) processes each sample from the EEG signal and predicts the value of the next sample \( (h_{t+1}) \) according to both the current sample \( (x_t) \) and the current fingerprint \( (f) \).

**Figure 3:** The recursive neural network used for generating EEG signal fingerprints.

Recursive neural networks are able to connect information from the previous sample to the current sample, allowing them to consider the structure of the EEG waves in each channel. However, interpreting EEGs requires considering not just the immediately adjacent signal information but also long distance signal patterns (e.g. alpha waves being interrupted by a sharp and slow wave complex or repeated bursts of high amplitude delta waves). In order to allow our EEG fingerprints to consider this type of long-distance information — that is, to consider the context of the entire signal when predicting the next sample, we adapt a special form of recursive neural network cell — the Long Short Term Memory\(^{16} \) (LSTM) cell — which are able to remember information for long periods of time.

The recursive neural network shown in Figure 3 is formally defined as follows. We define the parameter \( K \) as the fingerprint dimensionality, or the number of dimensions of the fingerprint vector such that \( K \ll N \times L \) (where \( N \) is the number of channels and \( L \) is the number of samples). This allows us to determine the fingerprint vector \( f \in \mathbb{R}^{1 \times K} \) for an EEG \( \mathbf{D} \in \mathbb{R}^{N \times L} \) by using the fingerprint as the internal memory state of the LSTM chain. In order to ensure that the fingerprint can be used to reconstruct portions of the EEG data, we define an additional parameter, \( W \), the sample window, which indicates the number of sub-EEG samples which should be predicted from each cell. This allows us to learn the optimal fingerprint for each document by determining the vector \( f \) which minimizes the cosine distance between the predicted values for each sample \( (h_t \ldots h_{t+W}) \) and the actual values \( (x_{t+1}, x_{t+2}, \ldots, x_{t+W+1}) \). Note, that the fingerprint vector and prediction vector are both \( K \)-dimensional, while each sample vector is only \( N \)-dimensional. Thus, before comparing, we must project the output vector \( h_t \) into \( N \) dimensions by defining a projection matrix \( \mathbf{W} \in \mathbb{R}^{K \times N} \) and a bias vector \( \mathbf{b} \in \mathbb{R}^{1 \times N} \). Unlike the fingerprint vectors, which are optimized for each individual EEG, \( \mathbf{W} \) and \( \mathbf{b} \) are optimized over the entire corpus. Thus, the optimal fingerprint vector \( f \) for each document was computed by minimizing the cosine distance between the output of the LSTM cell and the next \( W \) samples:
\[
f = \min_{f'} \sum_{i=0}^{L-W} \sum_{j=1}^{W} \cos(\text{lstm}(f', D^T) \cdot W + b, D^T)
\]

where \(\text{lstm}(f, D^T)\) refers to the standard LSTM loss function\(^{26}\). As defined, the recurrent neural network allows us to generate the optimal fingerprint \(f\) for each EEG signal by discovering the vector \(f\) which is best able to predict the progression of samples in each EEG recording according to the LSTM. In this way, the fingerprint is able to provide a substantially compressed view of the EEG signal while still retaining many of the long-term interactions and characteristics of the signal.

**Organizing EEG fingerprints in a similarity-based Hierarchy.** Rapid computation of similarity between EEG signals (or their fingerprints) is facilitated by the Fast Library for Approximate Nearest Neighbors\(^{25}\) (FLANN). FLANN provides implementations of a variety of highly-efficient structures for computing (or approximating) the nearest neighbors of vectors in high dimensions. This allowed us to not only compactly store the EEG signal information, but also to retrieve, for any EEG fingerprint, the most similar EEGs (i.e. the nearest neighbors) as measured by cosine distance. We used a k-means tree which allows for high precision retrieval of the nearest EEGs to any fingerprint by recursively clustering the fingerprint vectors using k-means clustering. The number of clusters is determined by FLANN’s auto-tuning mechanism.

**B. Query Analysis**

The purpose of query analysis is to identify the inclusion and exclusion criteria expressed in the query. For example, in the query “patients with shifting arrhythmic delta suspected of underlying cerebrovascular disease” two separate inclusion criteria are detected: “shifting arrhythmic delta” and “cerebrovascular disease”. Similarly, in the query “patients with dementia and no abnormal EEG”, there is one inclusion criterion, namely “dementia” and one exclusion criterion, namely “abnormal EEG”. To detect automatically the criteria, the following steps are used:

(Step 1) **Term Filtering:** Tokenization, lemmatization, and part-of-speech tagging using Stanford’s CoreNLP pipeline\(^7\) enables the filtering of terms that are not identified as a noun, verb, adverb, adjective, or preposition.

(Step 2) **Query Formulation:** Our approach for determining inclusion and exclusion criteria in the query relied on the same polarity and medical concept classifiers used (and previously described) for building the inverted index from EEG reports. Specifically, we considered two methods for recognizing inclusion and exclusion criteria: (a) phrase chunking using Stanford’s CoreNLP pipeline and (b) medical concept detection using the previously described classifier. In both cases, we relied on the previously described polarity classifier to distinguish between inclusion criteria (positive) and exclusion criteria (negative) based on the polarity of each phrase or concept.

(Step 3) **Query Expansion:** In order to account for the fact that many medical concepts can be expressed in multiple ways, we perform query expansion using the Unified Medical Language System (UMLS) to detect synonymous criteria. This is accomplished by expanding each criterion to include the set of all atoms in UMLS which have the same concept unique identifier (CUI) as the criteria. For example, “cerebrovascular disease” would be associated with 110 expansions, including “cerebral aneurysm”, “vascular ischemia”, “brain stem hemorrhage”, etc.

**C. Relevance Models**

The inclusion and exclusion criteria discerned from the query analysis were used by MERCuRY’s relevance models to assess the relevance of each EEG report against the given query. Two relevance models were considered (as illustrated in Figure 1): **Case 1**, which ignored the EEG signal fingerprints and **Case 2**, which incorporates them.

**Case 1.** This relevance model assigns a score to an individual EEG report based on the BM25F ranking function\(^{26}\). BM25F measures the relevance of an EEG report based on the frequency of mentions of each inclusion criterion and the absence of each exclusion criterion. Moreover, BM25F is capable of adjusting the score for each criterion based on the tiers in the posting list: that is, a criterion mention is scored according to both the polarity and the section of each mention. Formally, for an EEG report \(r\) and a query \(q = \{c_1, c_2, \ldots\}\) composed of individual inclusion and exclusion criteria \(c\), the BM25F relevance score is computed as:

\[
BM25F(r; q) = \sum_{c \in q} \frac{\tilde{x}_{r,c}}{K_1 + \tilde{x}_{r,c}} \cdot idf(c)
\]

where \(idf(c)\) is the inverse-document frequency of criterion \(c\) (i.e. the inverse of the number of documents mentioning \(c\)), \(K_1\) is a structuring parameter (in our case set to the standard\(^9\) value \(K_1 = 1.2\)) and \(\tilde{x}_{r,c}\) is a tier-normalizing criterion frequency measure. The tier-normalizing criterion frequency measure, \(\tilde{x}_{r,c}\), adjusts the frequency of criterion \(c\) in report \(r\) according to the polarity and section of each mention. Before defining this measure, we must account for
the fact that query analysis described above considers two ways of representing inclusion and exclusion criteria – (a) by phrases and (b) by typed medical concepts; thus, the tier-normalizing criterion frequency measure changes depending on which of these methods is used:

\[
\begin{align*}
\text{(a)} \quad \hat{x}_{r,c} &= \sum_{p,s} \frac{x_{r,c,s,p}}{(1+b)\left(\frac{t_{r,p,s}}{l_{p,s}}-1\right)} \\
\text{(b)} \quad \hat{x}_{r,c} &= \sum_{p,\ell} \frac{x_{r,c,\ell,p}}{(1+b)\left(\frac{t_{r,p,\ell}}{l_{p,\ell}}-1\right)}
\end{align*}
\]

(Case 1a) When an inclusion or exclusion criterion are expressed as a phrase, we defined \(\hat{x}_{r,c}\) (used by the BM25F function) in Equation (a), where \(x_{r,c,s,p}\) is the number of occurrences of criterion \(c\) with polarity \(p\) in section \(s\) of report \(r\); \(b\) is a normalizing parameter (in our case using the standard value \(b = 0.75\)), \(l_{r,p,s}\) is the number of terms with polarity \(p\) in section \(s\) of report \(r\), and \(l_{p,\ell}\) is the average number of terms with polarity \(p\) in section \(\ell\) across all reports.

(Case 1b) In this case, each criterion is represented as a medical concept. For this method, the tier-normalizing criterion frequency measure is restricted only to the sections pertinent to the type of medical concept. That is, medical problems, and medical tests are only searched in the HISTORY and CORRELATION sections; medical treatments are searched in the MEDICATIONS and CORRELATION sections; while EEG activities and EEG events are searched only in the DESCRIPTION and IMPRESSION sections. Consequently, the tier-normalizing criterion frequency measure (used in the BM25F function) is computed using Equation (b) where \(\ell\) indicates a section pertinent to the type of the medical concept used to express the criterion \(c\).

**Case 2.** The second relevance model considers both the information from EEG reports as well as the EEG signal fingerprints. It starts with the candidate patients discovered based on Case 1. The ranked list of patients is then updated based on the fingerprints associated with the most relevant patients’ EEGs. The rank updating procedure relies on two parameters: (1) \(\lambda\), the rank threshold parameter indicating how many of the initially retrieved patients should be used for re-ranking (in our experiments we set \(\lambda = 5\)), and (2) \(\delta\), the fingerprint selection parameter which determines the number of similar fingerprints to consider for each patient (in our experiments we set \(\delta = 3\)). The updated patient ranking is obtained as follows: for each patient \(p_x\) of the \(\lambda\)-highest ranked patients, we (i) find the fingerprint \(f_x\) associated with \(p_x\), (ii) use the hierarchy of EEG signal fingerprints (illustrated in Figure 2) from the multi-modal index to discover the \(\sigma\) most-similar fingerprints to \(f_x\), and (iii) insert the patients corresponding to these fingerprints into the ranked list of patients immediately after the patient \(p_x\), thus generating a new ranked list of patients.

**Evaluation**

We evaluated two aspects of the MERCuRY system: (1) the overall quality of patient cohorts discovered by the system and (2) the quality of the polarity classifier used to process the EEG reports and to detect exclusion criteria in queries.

**A. Evaluation of Patient Cohort Discovery**

We primarily evaluated the MERCuRY system according to its ability to retrieve patient cohorts. To this end, we asked three neurologists to generate a set of 5 evaluation queries each and then used them for evaluation. A sample of these queries is illustrated in Table 1. For each query, we retrieved the ten most relevant patients as well as a random sample of ten additional patients retrieved between ranks eleven and one hundred. We asked six relevance assessors to judge whether each of these patients belonged or did not belong to the given cohort. Moreover, the order of the documents (and queries) were randomized and judges were not told the ranked position of each patient. Each query and patient pair was judged by at least two relevance assessors, obtaining an inter-annotator agreement of 80.1% (measured by Cohen’s kappa).

<table>
<thead>
<tr>
<th>Patient Cohort Description (Queries)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. History of seizures and EEG with TIRDA without sharps, spikes, or electrographic seizures</td>
<td></td>
</tr>
<tr>
<td>2. History of Alzheimer dementia and normal EEG</td>
<td></td>
</tr>
<tr>
<td>3. Patients with altered mental status and EEG showing nonconvulsive status epilepticus (NCSE)</td>
<td></td>
</tr>
<tr>
<td>4. Patients under 18 years old with absence seizures</td>
<td></td>
</tr>
<tr>
<td>5. Patients over age 18 with history of developmental delay and EEG with electrographic seizures</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** Example queries used to evaluate the MERCuRY system

This experimental design allowed us to evaluate not only the set of patients retrieved for each cohort, but also the individual rank assigned to them. Specifically, we adopted standard measures for information retrieval effectiveness, where patients labeled as belonging to the cohort were considered relevant to the cohort query, and patients labelled as not belonging to the cohort were considered as non-relevant the cohort query. Because the relevance of a patient to a particular cohort can be difficult to automatically measure, we report multiple measures of retrieval quality.
Moreover, because our relevance assessments consider only a sample of the patients retrieved for each topic, we adopted two measures of ranked retrieval quality: the Mean Average Precision\(^{31}\) (MAP) and the Normalized Discounted Cumulative Gain\(^{31}\) (NDCG). The MAP provides a single measurement of the quality of patients retrieved at each rank for a particular topic. Likewise, the NDCG measures the gain in overall cohort quality obtained by including the patients retrieved at each rank. This gain is accumulated from the top-retrieved patient to the bottom-retrieved patient, with the gain of each patient discounted at lower ranks. Lastly, we computed the “Precision at 10” metric (P@10), which measures the ratio of patients retrieved in the first ranks which belong to the patient cohort. Although less statistically meaningful, the precision is the easiest to interpret in terms of clinical application in that a 100.00% Precision at 10 indicates that all of the patients returned above rank 10 completely satisfy all the criteria of the given cohort. By comparison, the other measures indicate the quality of the ranking produced by our system such that the MAP and NDCG scores capture the degree that a patient retrieved at each rank will more closely match the cohort criteria than patients retrieved at low ranks.

We measured the performance the MERCuRY system configured for the two relevance models illustrated in Figure 1: Case 1, in which only the EEG reports are considered and Case 2, in which both the EEG reports and EEG signal information is considered. In both cases, we considered both methods of representing inclusion and exclusion criteria: (a) using phrases composed of terms, and (b) using typed medical concepts. We compared these four combinations against three competitive baseline systems for text retrieval: Okapi BM25\(^{32}\) (BM25), language model retrieval using Dirichlet smoothing\(^{31}\) (LMD), and the Divergence from Randomness\(^{31}\) (DFR) framework using Poisson smoothing, Bernoulli and Zipfian normalization. Table 2 illustrates these results.

As shown, both configurations yield promising performance. Moreover, case 2 obtains the highest quality patient cohorts as measured by all three metrics. This shows that the multi-modal capabilities enabled by the EEG fingerprinting approach are able to identify patients who were missed when only the EEG reports were considered. The poor performance obtained by the baseline systems highlights the difficulty of automatically discovering patient cohorts. Moreover, the increase in performance obtained by MERCuRY Model 1 compared to Baseline 1 highlights the importance of medical language processing on EEG reports – particularly the role of the tiered index and the incorporation of exclusion spans. The highest performance was obtained by MERCuRY Model 2, showing the promise of including EEG signal information when discovering patient cohorts. This suggests that the content of EEG reports alone is not enough to adequately determine if a patient satisfies particular inclusion criteria. This finding is not surprising, as EEG reports were not written to completely replace the EEG signal, but rather to describe the important characteristics of the EEG recording which may be of interest to other neurologists. As such, EEG reports typically document only notable findings making it difficult to exclude patients based only on the text in the EEG reports. The superior performance obtained by MERCuRY Model 2 indicates that EEG fingerprinting is able to supplement the information in the EEG reports and bridge the gap between the high level description of EEG information in the text, and the low-level electrode potentials recorded in the EEG signal.

<table>
<thead>
<tr>
<th>Relevance Model</th>
<th>MAP</th>
<th>NDCG</th>
<th>P @ 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1: BM25</td>
<td>52.05%</td>
<td>66.41%</td>
<td>80.00%</td>
</tr>
<tr>
<td>Baseline 2: LMD</td>
<td>50.37%</td>
<td>65.90%</td>
<td>80.00%</td>
</tr>
<tr>
<td>Baseline 3: DFR</td>
<td>46.22%</td>
<td>59.35%</td>
<td>70.00%</td>
</tr>
<tr>
<td>MERCuRY: Case 1 (a)</td>
<td>58.59%</td>
<td>72.14%</td>
<td>90.00%</td>
</tr>
<tr>
<td>MERCuRY: Case 1 (b)</td>
<td>57.95%</td>
<td>70.34%</td>
<td>90.00%</td>
</tr>
<tr>
<td>MERCuRY: Case 2 (a)</td>
<td>70.43%</td>
<td>84.62%</td>
<td>100.00%</td>
</tr>
<tr>
<td>MERCuRY: Case 2 (b)</td>
<td>69.87%</td>
<td>83.21%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

**Table 2: Quality of patient cohorts**

B. Evaluation of Polarity Classification

We evaluated the quality of our automatic polarity detection approach by performing 10-fold cross validation on the 2,349 manual annotations we produced and measured precision, recall and \(F_1\) -measure, as shown in Table 3. We compared our classifier against two baseline classifiers, (a) “Baseline: Word Only” which uses only word features, and (b) “Baseline: UMLS Only” which uses only UMLS concept features. The MERCuRY classifier obtains substantially higher performance. Moreover, the poor performance of the baseline systems suggests that determining exclusion spans in text requires more information that lexical context and can be improved by incorporating NegEx and medical ontologies.

<table>
<thead>
<tr>
<th>Label</th>
<th>Precision</th>
<th>Recall</th>
<th>(F_1)-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: Word Only</td>
<td>24.50</td>
<td>29.65</td>
<td>59.35</td>
</tr>
<tr>
<td>Baseline: UMLS Only</td>
<td>37.08</td>
<td>14.22</td>
<td>20.55</td>
</tr>
<tr>
<td>MERCuRY</td>
<td>86.82</td>
<td>70.10</td>
<td>76.20</td>
</tr>
</tbody>
</table>

**Table 3: Polarity classification performance**
Discussion
In terms of polarity classification, the most common types of error were due to confusion regarding the exact boundary of negative regions of text, for example, the sentence “No focal or epileptiform features were identified in this record” was classified such that only “no focal” was negative, rather than the entire phrase “no focal or epileptiform features.” This indicates a failure by the incorporated standard natural language processing modules (part of speech and phrase chunking) to adapt to the clinical domain. One obvious path to improvement would be to annotate basic linguistic information on clinical documents – particularly on EEG reports.

Another common type of error was related to the binary granularity of polarity. For example, the excerpt there is a suggestion of a generalized spike and wave discharge in association with photic stimulation” the phrase “generalized spike and wave discharge” was labelled as having a negative polarity, despite the physician clearly indicating the possibility of such an activity. This implies that future work would be well supported by a more fine-grained approach to capturing the physicians’ beliefs, for example, by considering the assertions used in the 2010 i2b2 challenge. Unfortunately, introducing assertions requires overcoming additional barriers including increased risk of misclassification, and accounting for the degree of similarity between different assertion values.

In terms of patient cohort retrieval, it is clear that the two methods of representing inclusion and exclusion criteria – (a) using phrases of terms and (b) using typed concepts – do not provide any significant changes to cohort performance. Based on our analysis, we believe this is primarily due to the fact that there is little ambiguity in the types of concepts used in EEG reports: a particular phrase or term (e.g. heart attack) was always associated with the same concept type. Moreover, the types of concepts are almost completely restricted by the section they occur in (e.g. EEG activities and events do not occur in the HISTORY, MEDICATION, or CORRELATION sections). This suggests that considering EEG concepts alone does not provide any additional value to considering terms directly. Moreover, because the index records positional information, multi-term concepts (e.g. “slow and sharp wave”) are handled identically to multi-term phrases. Despite this, a number of errors were observed. First, neither phrase chunking nor concept detection is sufficient to fully capture the semantics of all inclusion criteria. For example, epileptiform activities were often described as attributes of a particular wave (e.g. “slow rhythmic delta [waves]”) where the individual concept (i.e. “delta [waves]”) is far less meaningful than its attributes (“slow” and “rhythmic”). This suggests that performance can be improved by not only accounting for the attributes of epileptiform activities but by adjusting the relevance model to ensure that mentions of attributes actually modify the correct term – that is, to ensure that “slow” actually modifies the same wave as “rhythmic”.

Finally, the substantial increase in performance when using the full multi-modal index shows that EEG fingerprints are able to recover relevant information omitted from EEG reports. Unfortunately, as the rank of retrieved patient’s decreases, the quality of the cohort obtained by finding similar patients using EEG fingerprints decreases. We investigated multiple values of $\lambda$ (the number of patients to be used for re-ranking) and $\sigma$ (the number of similar fingerprints to retrieve), and found that increasing these values can result in a decrease in performance. Regardless, we did observe that the fingerprints often identify patients which were not retrieved using the report text alone. Moreover, we believe that by further refining the fingerprints we can improve the quality of patients retrieved with higher values of $\lambda$.

Conclusion
In this paper we describe a patient cohort retrieval system that relies on a multi-modal multi-tiered index that organizes clinical information automatically processed from a big data resource of EEGs. Generating the index involved both medical language processing on EEG reports, but also a novel and highly efficient representation of the EEG signal recordings provided by a highly-performant Long Short Term Memory network. When evaluating the quality of patient cohorts obtained when considering both EEG reports and signal recordings, we have a Mean Average Precision of 70.43%. This high performance highlights the promise of multi-modal retrieval from text and signal data. The remaining barriers of high-accuracy patient cohort identification from EEGs that need to be removed will rely on: (1) incorporating a more fine-grained representation of inclusion and exclusion semantics discerned from EEG reports, (2) extending medical language processing for capturing spatial and temporal information, and (3) tightly correlating the information from EEG reports with the EEG signal recordings. In future work, we plan to address these barriers using recent developments in neural learning.

Acknowledgements
Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
References

Assessing patient and caregiver needs and challenges in information and symptom management: a study of primary brain tumors

Rebecca J. Hazen¹, Amanda Lazar, PhD², John H. Gennari, PhD¹
¹University of Washington, Seattle, WA; ²Northwestern University, Evanston, IL

Abstract
Brain cancer is a devastating diagnosis characterized by significant challenges and uncertainties for patients and their caregivers. Although mobile health and patient-facing technologies have been successfully implemented in many patient populations, tools and technologies to support these users are lacking. We conducted semi-structured interviews with 13 patients and caregivers, investigating experiences, challenges, interests, and preferences for managing symptoms and health information. We found that although current technology use in health-related activities was minimal, participants reported being highly willing to use such technologies to capture and manage information, provided they were designed according to the needs, interests, and abilities of these users. Participants felt that such tools could benefit patient care activities, and help to address information challenges for both current and future patients and caregivers. We present findings surrounding these challenges, behaviors, and motivations, and discuss considerations for the design of systems to support current and future patients and caregivers.

Introduction and Background
In 2016, nearly 78,000 US adults will be diagnosed with a primary brain tumor¹. Although rare, these tumors are a significant cause of morbidity and mortality, and are characterized by a high symptom burden, decreased health-related quality of life, and an often poor prognosis²-⁴. The delicate and complex structure and functions of the brain make these often aggressive tumors difficult to treat, with high recurrence rates contributing to poor long-term survival⁵. Despite advances in treatment technologies and approaches, nearly 42% of patients diagnosed with malignant disease will pass away within the first year after diagnosis¹. For these patients, overall 5-year survival rates fall at 34.4%, led by glioblastoma, the most common and aggressive high-grade tumor, which falls at 5.1%⁶.

In addition to poor prognosis, primary brain tumors are also associated with high symptom burden. As these tumors grow and invade the delicate tissues and spaces of the brain, patients experience neurological symptoms including headaches, seizures, issues with language, vision, memory, and hearing, as well as cognitive and motor impairments, and changes in behavior and personality⁵. The type and severity of these symptoms depend on the histology, location, and size of the tumor⁵. Treatments for these patients involves surgery to remove as much as the tumor as possible without damaging adjacent health tissues and functions, as well as radiation therapy and chemotherapy to kill any remaining tumor cells and prevent future growth or recurrence for as long as possible. Patients are often prescribed a variety of medications including steroids to reduce swelling in the brain, antiemetics to control nausea, analgesics for pain, and anti-seizure medications to provide symptom and side effect relief and prevent additional harm and neurological complications for the patient. Each of these medications and procedures bring their own risks, discomforts, and side effects, some of which may be lasting or even present years down the road. Because of these risks presented by these medications and procedures, and the high risk for recurrence associated with many primary brain tumors, the follow-up period of these patients typically lasts years, if not indefinitely.

Patients and their caregivers face many challenges as they take on new roles and responsibilities in navigating the diagnosis and treatment process. Many of these challenges center around the need for more data and information in the face of compounding uncertainties. In addition to taking an active role in decision-making activities, patients and their caregivers must work to manage complicated medication schedules and treatment protocols, endure complex neurological symptoms and toxic side effects, all while becoming experts in their own disease and care process.

With the increasingly pervasive nature of technology in daily life, mobile health and patient-facing technologies have been proposed and implemented as solutions to many challenging health issues including increasing medication adherence among individuals with HIV in Sub-Saharan Africa⁷, supporting self-management activities for patients with chronic diseases like diabetes⁸, and supporting personal health information management activities in breast cancer⁹-¹¹. Across health, reach of these technologies is great. A recent systematic review explored examples of self-tracking for health management in older adults, documenting the design and development of interventions across a wide range of health areas for these users¹². Additionally, there a multitude of applications are available to support mental health and individuals with mental illness, including PTSD Coach, an application
design by the Veterans Administration to support education, tracking, and management surrounding PTSD symptoms. Technology is also widely used in searching and accessing health information, with 25% of US adults acknowledging they use the internet for the purpose of self-diagnosis, and 72% of internet users reporting acknowledging using the internet to access or research health information in 2012. Electronic patient portals also provide a means for supporting communication, access, and health information management. Despite these impressive undertakings in the US and around the world, patient-facing tools and technologies to support individuals with rare conditions including primary brain tumors are still lacking.

Due to the nature of the disease and the potential for cognitive, physical, emotional, and behavioral impacts, caregivers often take on an active role in care, decision-making, and information seeking and management activities. In working to understand and address some of the challenges, several researchers explored the information needs of patients with brain tumors, as well as their caregivers. Although each of these studies provided a great deal of insight into different aspects of needs, challenges, and experiences for patients and caregivers, none sought to explore the use of health-related technologies in these activities in the context of designing future systems. In order to address this gap, and develop further understanding of the needs and challenges facing this population, we engaged patients and caregivers affected by primary brain tumors in semi-structured interviews. Through these interviews, we investigated symptoms and side effects, challenges, and current health tracking and information management approaches, as well as the use of technology in brain tumor related health activities. We also utilized brainstorming questions to generate ideas on how we as researchers, alongside participants serving as experts, might design tools to address some of these challenges, and better support future patients and caregivers.

Methods
Eligibility & Recruitment
Patients diagnosed with a primary brain tumor and primary caregivers of these patients were recruited to participate in semi-structured interviews coupled with a demographic, health, and technology use survey to explore needs, challenges, and experiences, as well as current and future use of health-related technologies. In addition to the eligibility criteria presented in Table 1, participants had to be able to understand the information presented during the consent process and make an informed decision regarding participation. In instances where cognitive impairment was a concern, participants were asked a series of questions during the consent process to ensure understanding of the purpose of the study and their rights as a participant. Patients could participate regardless of whether their caregivers chose to take part, and vice versa. Participants were recruited over a 9-month period through a local brain tumor support group as well as in the clinic at radiation therapy treatment centers in the Seattle area. University of Washington Institutional Review Board approval was obtained prior to commencing this research.

<table>
<thead>
<tr>
<th>Table 1: Eligibility information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patients:</strong></td>
</tr>
<tr>
<td>• Diagnosed and treated for a primary brain tumor within the past 5 years OR experienced a recurrence that required any form of treatment within the past 5 years</td>
</tr>
<tr>
<td>• Treatment involved some form of radiation therapy</td>
</tr>
<tr>
<td>• Able read, write, and speak English*</td>
</tr>
<tr>
<td>• At least 18 years of age</td>
</tr>
<tr>
<td><strong>Caregivers:</strong></td>
</tr>
<tr>
<td>• Primary caregiver of a patient meeting the patient eligibility criteria</td>
</tr>
<tr>
<td>• Able to read, write, and speak English</td>
</tr>
<tr>
<td>• At least 18 years of age</td>
</tr>
</tbody>
</table>

*This requirement was not used to exclude patients with aphasia or communication disorders, provided they were comfortable participating in the study, and could understand information presented and provide informed consent.

Data Collection
Demographic, Health, and Technology Use survey
Participants were asked to complete a survey to provide information surrounding demographics, diagnosis, and treatment history. The second component of this included questions regarding current use of technology in health related activities, with survey questions based on the 2012 Pew Health Tracking survey.
**Semi-Structured Interview**

Interviews were conducted as either 1-hour individual sessions (patient OR caregiver), or 2-hour patient-caregiver dyad sessions, according to participant preference. Individual interview sessions allowed patient and caregiver perspectives to be shared more freely and equally, and allowed for participation from individuals who did not have a patient or caregiver who was interested or able to participate (e.g. paid caregiver, severely impaired patient). Patient-caregiver combined sessions were offered both as a convenience, and as a way to allow for participation from individuals who may need extra support with language or memory, for example. Because many individuals travel long distances for care and then return home following the end of treatment, both in person and phone interview sessions were offered. Interviews were audio recorded and participants were compensated for their time.

**Data Analysis**

Interview data was transcribed and verified prior to data analysis. A thematic analysis was conducted with two separate coders analyzing the transcripts to identify codes and themes; each coder generated a codebook based on themes and codes. Codebooks were merged into a single codebook and additional codes were added and reconciled as additional transcripts were coded.

**Results**

A total of 13 participants (7 patients, 6 caregivers) were recruited to take part in this study, representing approximately 11 hours of interview data. Twelve individuals participated via in-person interviews, while one opted to do a phone interview due to location. Six participants opted for patient-caregiver dyad sessions, while one dyad participated separately, and 5 individuals participated independently. All interviews were conducted at the University of Washington Medical Center, however, many participants were treated outside of the UW Medicine system for part or all of their care. As such, information captured during these interviews was not limited to experiences at UWMC or its entities.

Demographic information for participants is presented in Table 2. Diagnoses included oligodendroglioma, anaplastic oligodendroglioma, oligodendroastrocytoma, astrocytoma, anaplastic astrocytoma, and glioblastoma, ranging from grades 2 to 4. Time since diagnosis ranged from 2 months to 4 years, 10 months, with one participant falling outside of the 5 year cut off, but having had a recurrence within the time frame of interest. All seven patient participants had undergone some form of radiation therapy, while six had also undergone surgery, and four had done chemotherapy. Four participants were currently in treatment, and three of the seven participants had reported recurrence or progression of disease. Participants represented a range of educational backgrounds with three participants having associate’s degrees, four with bachelor’s degrees, and six with graduate or professional degrees. Technology use information is discussed in findings section of this paper.

**Table 2: Demographic Information**

<table>
<thead>
<tr>
<th>Patients</th>
<th>Patients (P##)</th>
<th>Caregivers (C##)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female 4, Male 3</td>
<td>Female 5, Male 1</td>
</tr>
<tr>
<td>Age</td>
<td>Average 52.86, range 42-66</td>
<td>Average 50.3, range 39-63</td>
</tr>
<tr>
<td>Time since diagnosis</td>
<td>Average 20.2 months</td>
<td>N/A</td>
</tr>
<tr>
<td>Race</td>
<td>Caucasian 6, Not listed 1</td>
<td>Caucasian 5, Asian Indian 1</td>
</tr>
</tbody>
</table>

**Overall Findings**

Patients and caregivers faced a multitude of challenges as they worked to understand and manage complicated symptoms, side effects, medications, and treatment protocols. Participants looked to many different sources for information including clinicians, the internet, pamphlets and brochures, support groups, medication packaging information, scientific literature and clinical trials, patient advocacy groups, webinars, blogs and cancer forums, as well as trusted friends and family members to address different aspects of these challenges. Even with a wide range of information sources, many issues remained and in some situations, new challenges arose with increased access to information. Patients wanted to know what to expect in terms of symptoms and side effects, as well as what the future held for them. As a result of these experiences nearly all participants reported feeling lost, alone, scared, or overwhelmed at least once during the process. In this section, we present our findings, as outline in Table 3.
Throughout these interviews, participants reported experiencing over 60 different symptoms, side effects, and health events ranging from seizures, fatigue, nausea, anxiety and depression, to gross impairments in cognitive and motor functions, as well as changes in behavior and personality. While some were anticipated, others came as a surprise in terms of presence, severity, and duration. Many patients and their caregivers experienced challenges identifying and detecting these symptoms and side effects. Part of this challenge stemmed from the fact that symptoms can vary widely depending on the size and location of the tumor, and that they may be subtle in nature, or present slowly over time. A major contributing factor, however, was the fact that participants were not previously familiar with neurological symptoms. For example, five patients had experienced what they believed to be a stroke or dizzy spell, but was later diagnosed as a seizure. Additionally, one patient had been experiencing weakness on one side of their body during follow-up, but because they did not previously have any motor symptoms or impairments, they did not know to associate it with the tumor and report it to their clinicians. This later turned out to be a recurrence of the disease in another region of the brain. In several situations, caregivers noticed symptoms that patients were not aware of, mainly related to changes in cognition, behavior, and personality. In addition to challenges identifying and detecting symptoms, participants were also concerned with understanding the causes of the symptoms and side effects they were experiencing, and knowing how to manage and react to them. This involved trying to determine whether their experiences were related to the disease (i.e. progression or recurrence), or whether treatments or medications were responsible. Similarly, knowing what was normal or to be expected, when to be concerned, and what to do in response to these events was also a major challenge.

**Information Challenges: availability, depth, and presentation of information**

Having access to information was incredibly important to all participants across this study. There was a considerable amount of variation however, between participants in regards to overall satisfaction with the quality and quantity of information provided by their clinicians. In general, participants felt that they received more and better information during radiation treatment than while on chemotherapy at home. This was largely attributed to the fact that during radiation, patients were at the treatment center daily over a 6 to 8-week period, and met with clinicians once a week to discuss progress and questions. In contrast, patients on oral chemotherapy typically only saw clinicians once during each 6 to 8-week chemo cycle. The difference in quantity of face-to-face interactions, and the extended periods of time between appointments meant that patients and caregivers often waited longer to receive information, ask questions, and report concerns.

Participants also noted challenges involving the level of detail and presentation of information. Many participants stated that they would have appreciated more information surrounding diagnosis, treatment procedures, medications, side effects and potential complications, especially early on. P06 explained this in saying “I think [knowing more about] the medications [and] the treatments would have been helpful because you feel like you are jumping out of an airplane without a parachute when you start this journey.” For others, the amount and presentation of information was overwhelming, as described by P06 in saying “It’s actually kind of hard from my viewpoint, going

### Table 3. Interview Findings and Interpretations

<table>
<thead>
<tr>
<th>Themes</th>
<th>Findings and Interpretations</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current Challenges</strong></td>
<td>Symptom and side effect challenges</td>
<td>Patients and caregivers often struggle with unmet information needs and uncertainties, especially early in the diagnosis and treatment process</td>
</tr>
<tr>
<td></td>
<td>Information challenges</td>
<td></td>
</tr>
<tr>
<td><strong>Current Behaviors</strong></td>
<td>Tracking of health information</td>
<td>Current technology use is limited in these activities is minimal, as participants perceived many barriers and limitations; caregivers play an important role in many situations</td>
</tr>
<tr>
<td></td>
<td>Communication of health information</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caregiver role in management and care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technology use</td>
<td></td>
</tr>
<tr>
<td><strong>Future Behaviors and Motivations</strong></td>
<td>Self-tracking to support patient care and understanding</td>
<td>There is great potential for future technology design and development in this area, but barriers and needs of this population require careful consideration</td>
</tr>
<tr>
<td></td>
<td>Benefits of viewing previous patients data and experiences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Willingness and motivation</td>
<td></td>
</tr>
</tbody>
</table>
through this. I think they are talking at you a lot, and we're both pretty much in a little state of denial or something… You can’t keep track of all the information they are giving you and be able to register it enough to keep everything in your head.” Interestingly, C06, the other half of this patient-caregiver dyad, countered this, expressing frustrations over missing information several times, and summarizing in saying “Well, I would say that is true for you. I could keep track of everything they said because they didn’t tell me very much [both laughing]. I felt a lack of information, and you were overwhelmed by everything because of your state.” This feeling was most likely due to the fact that patients may not always be in a state cognitively or emotionally to take in and process the information being presented, especially following surgery and during treatment. At the same time, some caregivers felt they did not always receive the information they wanted or needed. One caregiver reported doing research on her own to learn the information and terminology so the clinicians would know and trust that she wanted more information.

Both patients and caregivers reported that they felt clinicians were often vague, unwilling, or unable to provide answers, especially related to prognosis. They acknowledged that at times, the information they wanted was not yet available, as was the case before biopsy to determine the type and grade of the tumor. In other cases, they attributed the hesitancy of clinicians to provide the level of information desired to the individualized nature of the disease and number of variables involved, as well as the lack of available clinical trial data for this small, rare disease population. Emotional considerations also came into play as participants believed that clinicians at times held back information or emphasized the positive extremes because they did not want to depress or upset the patients and their caregivers. This scenario was described by P04 in saying “Well I got the distinct impression that he was trying to invoke the power of positive thinking. He didn’t want to put any negative sort of doom-saying scenarios into the works because that can probably turn into a self-fulfilling prophecy.” Although this participant understood and appreciated the concern of their clinician, they went on to further describe that this information was also very important and necessary for setting expectations and making decisions about the future, as they often wanted to know whether they should be planning in terms of months or years.

Information Challenges: Applicability and the need for personalized information
It was nearly unanimously reported among patients and caregivers that they wanted to know what to expect throughout the treatment process and into the future, and that more and better information was necessary to make this happen. This included wanting to know what to expect for their individual situation in terms of symptoms, side effects, survival, quality of life, and the potential both long-term and short-term cognitive deficits. Participants frequently wanted to be able to compare their situations to information they found online or in the literature but experienced challenges in doing so. This was especially common when looking into information surrounding prognosis, where many factors can contribute to the likely outcome. Some participants were unable to find relevant studies or information at the time of decision-making, while others found that they were unlike the patients in the clinical trials due to factors including age, tumor size or location, and treatment history. Factors behind these challenges and frustrations extended beyond lack of availability in the literature, and into the clinic. Many patients reported repeatedly being told that every patient and every tumor is unique. Although likely intended to provide reassurance and discourage participants from reading into what they are finding in trials and on the internet, this type of communication lead to more confusion and uncertainty than relief. The frustrations associated with this type of communication were described by P07 in saying “What most of the doctors say is like… ‘oh well this is your tumor, and there is no other tumor like it. So your experience is your experience, and there’s no such thing as an average.’ And so they make these projections as to how I might or might not respond, but they don’t know, and they always quantify it saying ‘I can’t tell you because it’s you and your tumor, and it’s not somebody with their tumor that’s had the experience that’s in the statistics.’ And so the trouble with that is you come away without any knowledge whatsoever…” Concerns and the desire for in-depth, personalized and relatable information did not end once treatment was over. For certain participants, setting clear expectations relieved stress, and helped them prepare for and accept future possibilities.

Information Challenges: Credibility, cognitive impairment, and emotions
Patients and caregivers often looked to others for information, guidance, and support throughout the journey. As most participants did not know anyone else who had a brain tumor prior to their own diagnosis, they were often forced to look to the internet for information to learn about patient and caregiver experiences by way of blogs, personal websites, and online forums. Several participants felt that without being able to interact with the source of the information and assess their cognitive and emotional state, they could not be certain of the credibility of the information being shared. This concern was linked to the fact that changes in personality, behavior, and cognition are common in individuals with brain tumors, thus, there is increased potential for sharing of misinformation,
whether intentional or not. Interestingly, this sense of distrust and skepticism was more common amongst patient participants, whereas several caregivers reported more positive impressions, likely due to the fact that they were interested in learning different types of information from these sources. There were also great concern involving the internet, both in terms of the potential for discovering information that is scary or upsetting, as well as finding incorrect information. Most participants reported using the internet to look up information in the early stages of the disease and treatment process, but had since significantly decreased or completely stopped to avoid the stress and reduce the potential for misinformation.

**Current Behaviors: Tracking, Communication, Caregivers and Technology Use in Information Management**

*Tracking and recording health information*

Keeping track of health, symptom, and side effect information can be a valuable component of health information management. The majority of current tracking activities were informal in nature with participants largely relying on memory, or taking notes in journals or on pieces of paper. Participants primarily noted when they noticed changes in a certain symptom or side effect, or if something unexpected occurred that they wanted to bring up with their care team. For the majority of these participants, recording of information related to symptoms, side effects, and experiences was not intended to serve as a written record, but was used instead to support memory, communication, and organization of information. Motivations and reasons for tracking or taking notes varied. One caregiver was mentally tracking concerning symptoms for the patient in order to help identify the cause, and hopefully find some sort of solution or resolution. For another participant, tracking of seizure information led to the discovery of a correlation between the frequency of her seizures and her menstrual cycle. For the participant who was sharing caregiving responsibilities with another family member, tracking came about as a byproduct of communication and comparing notes rather than a deliberate decision to track or record information.

Reasons why formal tracking or recording health related information did not take place, using mobile health applications or otherwise, typically fell into three categories: (1) participants were not explicitly asked to capture data by their clinicians, (2) participants felt as though they were seeing the care team frequently enough that symptoms and side effects were being adequately assessed, or (3) participants felt that they did not know what or how to track health information. Category two was especially interesting with this population as this lead to the discovery that tracking and reporting needs and behaviors varied significantly depending on the stage in the treatment or follow-up process. As previously mentioned, patients are seen far more frequently when undergoing radiation therapy, as compared to at home oral chemotherapy. Two participants explicitly stated that they felt they didn’t need to track or record information related to their symptoms during this time period because the clinicians were already keeping track of the information that they were interested in. In contrast, during oral chemotherapy, patients were seen far less frequently, and often felt like they didn’t know how or what to track.

*Communication of Health Related Information*

Communication activities, preferences, and behaviors varied among participants as well. Email was by far the preferred method for contacting clinicians with routine questions and prescription refill requests. Several participants appreciated when clinicians were willing to use email, avoiding the inevitable game of phone tag that typically ensued when patients called in with questions. Despite being convenient and familiar, participants also experienced some issues with using email. Patients and caregiver feared overlooking important symptoms, but also did not want to burden healthcare providers with unnecessary or ‘stupid’ questions, despite their need for information. P07 described a situation where they had attempted to reach out to a clinician via email, but was conflicted and worried about being a bother. Others also reported that they had reached out to certain clinicians via email, but abandoned it as a method of communication after failing to receive a response. Clinician preferences for communication during treatment and follow-up, especially regarding email, were rarely made clear, thus, knowing who to contact, and how to reach them when help was needed remained a challenge. The use of secure messaging through the patient portal was used much less frequently, and was seen as much less convenient than email.

*Caregiver roles, responsibilities, and challenges in care and information management*

In several cases, patients were either unable, or unwilling, to participate in managing and capturing this information during parts or all of their care process due to the cognitive, emotional, and physical impact of the disease and treatment process. This meant that caregivers took on much of the responsibility for managing information provided by clinicians, accessing test results through the patient portal, transferring records between healthcare organizations, scheduling appointments, researching treatment options, medications, diets and nutrition, refilling prescriptions, dealing with insurance and managing financial and logistical aspects of the care process. At the same time,
caregivers faced unique challenges in that their questions and concerns often did not take priority, and at times went unaddressed despite the fact that they were often heavily involved in research, decision-making, and patient care activities. In several cases, caregivers felt as though they could not ask sensitive questions in front of the patient, or patients felt that their questions were more of a priority given the limited amount of time they had with the clinicians. In other instances, patients were feeling well enough to take on all of the information management activities, and didn’t feel that they needed caregiver assistance or participation in appointments, and may have blocked caregiver access to clinicians, despite the fact that they may have had their own questions and concerns. Additionally, cognitive impairments and issues with behavior and personality led to at least one patient becoming confrontational, and not wanting caregivers to talk about symptoms or ask questions during appointments. In three cases, participants felt that separate caregiver appointments might have been beneficial to getting caregiver questions answered. Another aspect of information management that arose was the issue of co-caregivers. Because caregiving is often time consuming and demanding, some caregivers shared responsibilities with a sibling or spouse. This meant that multiple people were involved in activities like scheduling appointments and managing medications, as well as capturing information about the patient’s condition, and comparing notes over time.

**Technology Use**

Despite the fact that all participants reported using technology in their daily lives, with 12 of the 13 participants owning and using a smartphone regularly, the use of personal technologies in health activities related to their cancer diagnosis was very limited. From the technology use survey, we found that all 6 caregiver participants reported helping to keep tracking of information for the patient, with 4 relying on memory, and 2 using paper. For the patients tracking their own health information, three relied on memory, and two used paper. Additionally, two patient participants reported using smartphone applications for other aspects of health and well-being, unrelated to their cancer treatment. A few participants mentioned using electronic calendars for managing appointment information, as well as spreadsheets for keeping track of appointment schedules, medications, and financial information. In two cases, email was used as a means of communication as well as a method for documentation and management of information. One caregiver had tried using a medication reminder application for the patient they cared for, but found that it was difficult to maintain, and easy to ignore. It was not worth the effort of making updates to reflect frequent changes in medication type, dosage, and frequency, especially as the patient was not responding to the alerts.

Several participants had used smartphone applications for tracking exercise and fitness, diet, menstrual cycles, meditation, and information related to other health conditions. Despite everyday use of technology, paper was largely considered to be the easiest, fastest, and most convenient option. For several participants, paper was more likely to be on hand and immediately accessible compared to cellphones, computers, or tablets. Other participants felt that they could not navigate these devices quickly or efficiently enough to record the information of interest. With paper, participants could jot down questions or notes whenever they came to them without having to go find a device, log in, and access a feature for recording information. In the clinic, participants could write out their questions beforehand, and quickly record responses alongside those questions during the appointment. Another benefit of paper, for one participant, was that they could practice cognitive and motor skills including hand-eye coordination when writing in their notebook. For these participants, limited technology use in health was not related to distrust, but convenience was a matter of functionality and convenience.

Patient portal usage was also explored in this study. The majority of participants had accessed their own patient portal at least once, or the portal of the patient they were providing care for. They felt that in general, the features and information was useful, but often limited. They reported accessing the portal to look at blood work, radiology reports, and other test results, and occasionally to verify appointment information. There were also challenges and limitations associated with patient portals, however. Access was an issue for several participants, as they had experienced difficulties with the system failures or had forgotten their log in information. For some participants, patient portal access was not offered until later in their treatment process, and historical information was not available through the system. Additionally, many caregivers did not know whether they could receive access to the portal of the patient they provided care for, and instead had to rely on having the patient log in and allow them to view the information. Some patients were receiving care from providers across different healthcare systems, or at healthcare systems without integrated portals so information was inaccessible or dispersed across portal systems. They also expressed frustrations related to missing data and the often limited nature of the information provided in reports. Participants felt that they should have immediate access to their own health data, and that delays in posting this information was unacceptable. In some cases, the patient and their caregiver already knew all of the information.
by the time it was posted to the portal. Other noted challenges and frustrations involving systems not being user friendly. Many of these challenges and limitations led to users to abandon the portals, instead relying on other means of communication or information access such as requesting print outs in during clinic visits. Several participants noted that these portals could be much more useful if they provided additional information, features, and functionalities including the ability to record their own information and notes, and access to information and resources about the patient’s diagnosis and resources would be helpful to have.

**Future Behaviors and Motivations**

*Self-tracking to support patient care and understanding*

Many participants felt that technology could potentially be beneficial for supporting tracking and managing health information, provided that it was designed to meet their needs, interests, and abilities. They felt that having a more complete record of information for clinicians to work through in making decisions and determining causes of symptoms and side effects could be helpful. Additionally, some felt tracking could be helpful in understanding and finding correlations in their own data, and could reduce the need for memory in recalling questions and information related to experiences in the days and weeks between appointments. However, because participants were not currently using any sorts or applications or tools in tracking or managing their health information, and because the only reference they had was the limited functionality of current patient portals, there was some reluctance to say with confidence that they would switch over if tools were to become available. Some of this was resistant to change as they had already developed methods that worked well enough for them, relying on paper and memory. As such, current behaviors and motivations, as well as the needs, interests, and abilities of these patients and their caregivers would need to be carefully considered in future technology design.

*Benefits for future patients and caregivers*

Thinking back on the challenges they faced, many participants felt that they could have benefitted greatly from the data and experiences of previous patients and caregivers, especially in helping to know what to expect, and to determine whether what they were experiencing was normal. One caregiver felt having access to such data, through the mediation and interpretation of clinicians, would have helped them to ensure that they were not being naïve in their expectations, while another felt that this would help them in reassuring and supporting the patient they were caring for, especially early on. Data of interest included quality of life and survival, and information surrounding type, severity, and duration of symptoms and side effects, as well as whether they resolved on their own or if intervention was necessary and effective.

Another area where this would have been useful is in working through information surrounding prognosis and what to expect as the disease progresses. As clinical trial data is incredibly limited for patients with primary brain tumors, clinicians are often forced to rely on their own experiences and anecdotal evidence when providing this information. Several participants felt having access to more complete tracked data could potentially help clinicians in providing more confident estimates of prognosis. Many participants were highly motivated by data and felt that having the option of knowing the potential possibilities, and having statistics to back that up would be of great comfort and benefit to themselves as well as future patients and caregivers. Although the majority of participants would have appreciated seeing such data or experience information, others had strong preferences about what they did and did not want to see. One participant indicated that they were not at all interested in seeing or hearing about experiences of previous patients with their condition because of the personal nature, and the eventual uncomfortable reality that it is a terminal disease. The same participant was comfortable with viewing data, but reaffirmed that they were not interested in anything related to personal experiences. They felt that the experience information may have been helpful for their caregiver, however.

*Willingness and motivation for Future Tracking*

Recognizing that patient and caregiver tracking and reporting of these data and experiences would be essential to having it available to address these challenges for future patients and caregivers, these participants were almost entirely unanimous in their willingness to provide that data. Some participants like P07 felt that this would be a good way to contribute, saying “I said right from the beginning, I would be happy to help down the road… I am not the last one that’s going to get this diagnosis, there’s people coming up all the time with it. If I can help somebody else, I would be happy to do that.” For others, the major motivation was mainly in having the data available for current patient care activities and psychological comfort.

**Discussion**
Some of the significant findings from this study was the participants strong desire for more information, a willingness to contribute to their own care and that of future patients and caregivers, and the desire for a better understanding of their current situation and potential future with this disease. Due to the number of variables involved and problematic uncertainty, it is currently impossible for clinicians to provide detailed estimates of prognosis for the majority of these participants, which is a major point of confusion and frustration. Optimism and positivity from clinicians was appreciated at times, but also detracted from overall understanding. Although likely intended to inspire optimism to ‘defy the median survival rates,’ the approach and telling patients that they are unique, and not like other patients was not helpful. Rather, participants felt that having access to actual data through clinicians, and to compare themselves to know whether their experiences were as expected, and to identify possible trajectories was incredibly empowering. They felt data could help patients, caregivers, and clinicians to have these honest conversations, even if it meant that there were still many possible trajectories and outcomes. Also, acknowledging that this data likely does not yet exist motivated these participants to be an active part of the data collection process, providing benefit beyond their own patient care.

In addition to these findings, we identified three major considerations for future design and development in this space. First, when discussing the design of technologies for users with complex medical conditions, the interests, needs, and abilities of the intended users require special attention. This is not unique to brain tumors, but because the brain controls so many functions, and these tumors can impact a wide variety of abilities, incorporating this information into design discussions from the start is important. For individuals with potential motor impairments or challenges with language, requirements for multiple methods of data entry (e.g. text, speech, pick list) may be necessary. Minimizing the need for memory, and streamlining the process to avoid redundancies was important. Participants felt that the technology should be ‘smart’ enough to pull relevant data directly from the medical record with minimal efforts from the patients and caregivers. They also emphasized that methods of data collection would be important for assessing reliability, preferring subjective measures over long narratives. Additionally, processes for verifying information would also be beneficial. Ultimately, data entry has to be convenient or else it will not occur. Considerations for shared access for caregivers was also necessary as some patients may be unwilling or unable to participate in tracking activities during part or all of their treatment process.

Second, because participants were satisfied with their current methods of tracking and managing information, there will likely be a barrier to overcome in incorporating technologies into their routines. For these participants, the potential benefits of having a more complete record to support discussions, and the potential that this data might help patients to support their own discovery process was valuable. Integrating the use of this data into clinical encounters might be one way of overcoming that barrier. Incorporation into the clinical encounter also provides an opportunity for clinicians to review the information with patients and their caregivers to ensure reliability and completeness prior to submitting it to a repository. Patients and caregivers were more trusting of the information that they were given when it came directly from or was verified by their clinicians. Although data integrity can be an issue with many types of patient-reported data, it is especially a concern here where cognitive, emotional, and behavioral factors may influence the data, and where the amount of data available is likely to be very small. Another benefit of integration of these applications and discussions into the clinical encounter is that patients may feel like they have the tools to help lead discussions surrounding likely causes or correlations between side effects and medications. Clinicians can also control the conversation and interpretation of data, and can guide conversations based on an understanding of patient values and preferences. As a final consideration, participants felt that any solution that is put into place needs to be easier and more convenient than pen and paper, which for these participants, would be very difficult to achieve. To overcome this, there must be significant motivation and benefit for them. For these participants, some saw benefit for patient care activities, while others were incredibly motivated by the chance to provide data and give back to future patients and caregivers. This may be generalizable to other patient populations where there are few patients, and many uncertainties.

There are two main limitations to acknowledge in considering the findings of this study. First, this study involved a relatively small number of participants due to challenges with recruitment, the rare nature of primary brain tumors, and the severe impact and burden they place on patients and their caregivers. Several potential participants were unable to participate due to demanding treatment schedules and the lasting effects of the disease. Additionally, these participants may not be truly representative of the overall brain tumor patient population, especially in terms of education and race. This was largely due to the fact that the patient population is small to begin with, and we only recruited within the Seattle area. Several of these participants however traveled from surrounding areas and states for specialized treatment at facilities in the Seattle area.
Mobile-health and patient-facing technologies are not yet readily available for this population, however, the potential for tool design in this area is vast. Not only would a smartly designed tool, created to overcome barriers and designed with the needs and abilities of these particular users in mind, support current patients and their caregivers, but also those who will be stricken with this disease in the future. These patients and caregivers face numerous challenges and uncertainties, for which this data could provide a tremendous amount of insight and relief. Designing a patient facing tool for this area of healthcare has great potential to benefit not only this subset of patients, but also provide insight into the overwhelming number of little understood diseases.

References:
Model Guided Design and Development Process for an Electronic Health Record Training Program

Ze He, Ph.D. 1,2, Jenna Marquard, Ph.D. 1, Elizabeth Henneman, Ph.D. 3
1. Department of Mechanical and Industrial Engineering, University of Massachusetts Amherst, Amherst, MA
2. College of Computer and Information Sciences, University of Massachusetts Amherst, Amherst, MA
3. College of Nursing, University of Massachusetts Amherst, Amherst, MA

Abstract
Effective user training is important to ensure electronic health record (EHR) implementation success. Though many previous studies report best practice principles and success and failure stories, current EHR training is largely empirically-based and often lacks theoretical guidance. In addition, the process of training development is underemphasized and underreported. A white paper by the American Medical Informatics Association called for models of user training for clinical information system implementation; existing instructional development models from learning theory provide a basis to meet this call. We describe in this paper our experiences and lessons learned as we adapted several instructional development models to guide our development of EHR user training. Specifically, we focus on two key aspects of this training development: training content and training process.

Introduction
Effective user training has been recognized as an important factor to improve health information technology (IT) implementation success, but development of a successful training program remains a challenging task. Part of the reason is that health care organizations lack industry-wide best practices in electronic health record (EHR) training, so often have to learn from their organization’s own training experiences, with can be costly and inefficient. While current literature often introduces some design principles, there is still a gap in guidance on how to translate those principles into varied clinical practices or settings. This literature may report on how organizations conducted the training and general principles, but the detailed process of developing training programs is often underemphasized and underreported. The literature may also include lessons from costly “do-overs” after the training program go-lives. Ideally, by following an instructional development model, those issues might be detected before the go-live, which could save money and time for clinicians and health care organizations.

American Medical Informatics Association (AMIA) board members published a white paper entitled “Core Content for the Subspecialty of Clinical Informatics”, which identifies core educational content for clinical informatics. With continued national mandates for health IT adoption and use, the training needs for information systems in clinical settings, most notably EHRs, are expected to grow. Among the numerous knowledge and skill content areas addressed by the AMIA board members, knowledge under “clinical information system implementation” includes “models of user training and support processes that can meet clinician needs” as one of three sub-topics.

Instructional development is a well-established domain, with over 100 validated instructional development models, and some of them may be particularly applicable to EHR training. For example, the Instructional Systems Development (ISD) model is a generic model developed in the 1950s to meet the needs of military and commercial organizations. In ISD, training consists of five basic stages: analysis (A), design (D), development (D), implementation (I), and evaluation (E); referred to as the ADDIE model. The ADDIE model is comprehensive in content and systematic in procedure, and covers almost all necessary components needed to carry out an effective training program. To meet a variety of specific training needs, other models build on the ISD model, and improve some of these stages. For example, the development stage may take a long period of time, and may fail to meet the pace of technology updates, e.g., EHR. The rapid prototyping model tries to facilitate training development by performing several elements...
simultaneously, condensing the generic ADDIE model into a four-level process, including: a) performing a needs analysis, b) constructing a prototype, c) utilizing the prototype to perform research, d) implementing the final system. The ISD model also fails to embrace the fact that reality can be different from what was planned, and may change over time, which is typical in complicated health care settings. Therefore, the R2D2 model (Recursive, reflective, design and development) provides the ability to update the training plan over time. Reflection involves critically considering work to date, and revising training plans and materials based on observation and other feedback. The recursive nature of the process means making the same decisions many times throughout the design and development process, so initial decisions are not necessarily the “final ones”. The implication of this model is that a training design team should actively reflect on and revise current training to meet ever-changing situations.

The current accounts for successes or failures in current EHR training often focus on the final format of training, and it is rare to see detailed descriptions of the training development processes. In this paper, we try to bridge that gap by discussing our experiences developing an EHR training program. The primary focus of this paper is to detail the systematic process of developing the training and lessons learned through our iterative development process. We consider training development as a problem that should take a sociotechnical system perspective, and adopted an iterative approach similar to rapid prototyping model and R2D2. We also conceptualize a two-phased EHR training development focused on training content and training process. We discuss what we learned during the design and development process that led to a successful training session.

**Methods**

Instructional models typically suggest that before training program development, one needs to conduct a thorough analysis to define training needs. We first present how to conduct this type of detailed analysis from a sociotechnical systems perspective, and then detail how we designed and developed training content and processes based on these analysis results. Figure 1 outlines the conceptual model of training development we adapted from ISD, R2D2, and rapid prototyping for EHR training program development. We use the model to guide the design and development of EHR training for nursing students at a large public university, and we address how we conducted each stage of this model in this section, with focus on the development of content and process. We chose to use the clinical process of obtaining a medication history prior to the design phase, because it is an error-prone process involving comprehensive EHR-related competencies (e.g., information retrieval, documentation, decision making), and it has been used in comprehensive practice sessions in previous EHR implementation training in other health organizations.

![Figure 1. Conceptual model of training design and development process](image)

The authors of the AMIA white paper on “Core Content for the Subspecialty of Clinical Informatics” adopt a holistic sociotechnical systems perspective, depicting clinical informatics at the intersection of three domains: clinical care, information technology, and health systems. In order to develop a training program that fits the need of clinicians and reflects the complex nature of healthcare settings, it is necessary to...
consider EHR training as a sociotechnical systems problem. This view is particularly important during the analysis stage of an EHR training needs assessment, because these analysis results will influence decisions related to training design and delivery.

The training analysis phase should define training goals, and gain an in-depth understanding about the users and their required tasks, including their typical clinical practices, informatics needs, responsibilities, and current competencies. This information can be used to determine what resources, including EHR functions, physical environment, and infrastructure will best serve users’ needs.

**Analysis of training goals**

The goal of our training program is to help senior nursing students use an EHR efficiently and accurately while obtaining a patient’s medication history. More specifically, after training, nursing students should be able to:

- Use functions related to medication history taking to navigate, search for information, retrieve information, and document new information into EHR
- Reconcile any potential discrepancies in patient information in the system, such as omissions and out-of-date medications

**Analysis of stakeholders**

We involved a multi-disciplinary team in the training design process, including individuals with nursing, computer science, and human factors backgrounds. In addition, the targeted trainee group participated extensively in the development process.

**Analysis of trainees’ characteristics**

Training analysis must address the trainee’s current competencies in completing the clinical process of interest, and current EHR competencies, in order to define their training needs. If the trainee is insufficiently prepared to complete the clinical care process, training must first address that competency prior to conducting EHR-specific training. Studies reported best practice principles in EHR implementation training that align with this approach, which identified a need to “assess users’ skills and training needs” and “match training to users’ needs” as important best practices. The analysis must also assess any technical competency prerequisites for using the EHR (e.g., using a keyboard and mouse, opening software).

In terms of competencies, the nursing students in our program had received classroom training in pharmacology and obtaining a medication history, had some clinical experience with EHRs through their clinical experiences (though not the EHRs used during training), and were competent with basic operations with computers (e.g., using a mouse and keyboard, and opening software).

**Analysis of clinical tasks**

To determine what tasks to train users on, an analysis should start with a clinical care process of interest, rather than the technical capabilities or functionalities of the EHR. The analysis should also determine the trainee’s roles and responsibilities within the process. In our training analysis, we first analyzed the clinical process carried out by nurses when obtaining a medication history. We built on an existing clinical process flowchart defining required clinical tasks.

Starting training from the perspective of a familiar clinical process is consistent with adult learning theory, which states that adults are more motivated to learn things relevant to their needs and experience. This analysis can help avoid overloading the trainee with irrelevant information and EHR functionalities, which is a typical cause to ineffective training.

**Analysis of EHR technical functions**

The training analysis should determine what technical aspects of the EHR (e.g., functions) are used during the clinical process of interest. After extensive review of the literature and analysis of EHRs, we identified what EHR functions were required to complete the clinical process, including both overhead functions (e.g., logging into the system) and domain functions (e.g., reviewing the current medication list in the
EHR) as described by Zhang et al. Specifically, trainees need at least six EHR functions to complete the clinical task, including:

1. Login to the system
2. Search for and retrieve a specific patient’s chart
3. Retrieve basic patient information, such as allergy, identity, vitals
4. Retrieve and document a patient’s medication list
5. Retrieve and navigate through a patient’s problem list
6. Retrieve and navigate through a patient’s past history, such as social, medication, surgery

**Analysis of task-EHR integration/ workflow**

Sociotechnical systems models suggest that we need to analyze the often-unexpected impact of an EHR on existing system elements, such as workflow. For example, EHRs can impose high cognitive load for clinicians, making clinical tasks error-prone. Therefore we clearly mapped out the links between the clinical care process and the EHR use, including what tasks should be completed with the EHR, when in the process the EHR will be used, and how the technology will impact the trainee’s workflow for the clinical process of interest. The results from this analysis are used in a clinical task-EHR integration/work training section. As an EHR will likely impact the clinical care process, the analysis helps mitigate negative impacts of these changes, and explicitly address any redesigned clinical workflows in the training.

**Analysis of resources and constraints**

Based on nursing students’ typical schedule, an ideal training session should be no more than 90 minutes to achieve best attendance rate and keep trainees attention. The physical environment is a Class Lab in nursing school, with a projector and white board. The training utilized laptop PCs with Windows Operating Systems with each trainee using their individual computer. The EHR is web based, which is accessible using a web browser.

**Design and development**

Based on the analysis results, we needed to make two main decisions: what to train (content) and how to train (process). We first decided necessary sections in the training program, which include clinical process review, pre-training evaluation of EHR competency, EHR use training, workflow integration training, and post-training evaluation. Then we worked on content and process for each section. Before we involved students in the study, we obtained approval from the Institutional Review Board.

Following the idea from the instruction development models, the general strategy to develop training content was to prototype materials for each training section, then test the materials with real trainees as they used the materials and processes. We worked to understand their perceptions, and reveal any confusions and ambiguity in the training materials. We worked with six different trainees individually on materials and processes, until we could identify no issues to be addressed and we were ready to roll out the training.

With the first three trainees, we focused on the content of the training. We asked each trainee to go through the training materials in order, observed them learning without imposing any time limit, and documented their time and questions. We asked them to think out loud so we could understand points of confusion, and tried to clarify verbally until they understood. With the second three trainees, we focused on the process of the training. We followed specific training protocols, with time limits for each section. We observed and documented issues as trainees went through the training, and communicated with them about the issues and their concerns at the end of the training.

We summarize major issues identified with each participant, and our corresponding changes in Table 1.
<table>
<thead>
<tr>
<th>Participant</th>
<th>Major Issues</th>
<th>Training Design Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a. The trainee was unfamiliar with some EHR technical terms when the terms are not commonly used in clinical care.</td>
<td>a. Simplify or define terms to fit clinicians’ language. For example, “search for information” sounds more familiar than “information retrieval”, and “enter data/notes” may be more familiar than “documentation”.</td>
</tr>
<tr>
<td></td>
<td>b. The trainee lost attention when a video tutorial introduced too much information at a time</td>
<td>b. Instead of introducing all EHR functions at once, introduce two functions at a time, followed by a hands-on practice</td>
</tr>
<tr>
<td></td>
<td>c. The trainee remained confused about what she needed to do with EHR with different medication history sources after clinical process review</td>
<td>c. Add a review question about reconciling with multiple medication history sources and using EHR, and provide answers</td>
</tr>
<tr>
<td></td>
<td>d. The trainee did not understand why the EHR-workflow tutorial was related to “how they conduct clinical care”, and did not understand the task-EHR flowchart mapping</td>
<td>d. Redevelop task-EHR integration tutorial, making it more relevant to clinical care, and present the integrated workflow chart in resemblance to clinical process diagram</td>
</tr>
<tr>
<td></td>
<td>e. The trainee did not feel comfortable in the physical environment, feeling too hot</td>
<td>e. Conduct training in an air-conditioned environment</td>
</tr>
<tr>
<td>2</td>
<td>a. The trainee did not get some key details of instructions printed on paper</td>
<td>a. Add a training instruction transcript for all sections, and verbally communicate what to do before each section</td>
</tr>
<tr>
<td></td>
<td>b. The trainee did not know what to do during hand-on exercise besides repeating what was demonstrated in the video</td>
<td>b. Add specific goals and assignments to the hands-on exercise, which covers the contents just demonstrated</td>
</tr>
<tr>
<td>3</td>
<td>a. The trainee had a problem organizing and finding files</td>
<td>a. Distribute files only when they are needed for each section, and ask trainees to put away in a file folder in checklist</td>
</tr>
<tr>
<td></td>
<td>b. The trainee frequently asked “what’s next”</td>
<td>b. Add a file called progress checklist to list all training sections, tasks and time in order, so a trainee can have expectations and more control</td>
</tr>
<tr>
<td>4</td>
<td>a. The trainee was quicker to complete tasks than the time set for the training, and had to wait for next sections to begin</td>
<td>a. In the instruction/protocol, the time set for each section is changed to the maximum time allowance, making it more self-paced; a smaller learning group (1 to 3 trainees per training session) can better accommodate different learning capabilities.</td>
</tr>
<tr>
<td></td>
<td>b. The trainee’s cell phone rang during the training</td>
<td>b. Add a welcome PowerPoint slide projected to whiteboard with silenced cell phone reminder</td>
</tr>
<tr>
<td>5</td>
<td>The trainee seemed to miss some key information about task-EHR integration tutorial printed on paper</td>
<td>Provide video version tutorial in addition to paper version</td>
</tr>
<tr>
<td>6</td>
<td>The trainee went through the materials and processes very smoothly. Ready for go-live.</td>
<td>None.</td>
</tr>
</tbody>
</table>
After 6 iterations of development, we finalized the training program as described in Table 2.

**Table 2.** Training program outline

<table>
<thead>
<tr>
<th>Section</th>
<th>Brief Description of Process</th>
<th>Duration (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check in</td>
<td>Trainees are seated and learn the goals and process of the training</td>
<td>5</td>
</tr>
<tr>
<td>Clinical process review</td>
<td>Trainees review the medication history taking process, answer review question, and get feedback via presentation of the solution</td>
<td>5</td>
</tr>
<tr>
<td>EHR Case 1</td>
<td>Trainees login to the EHR and complete patient Case 1 with no training, as a pre-test.</td>
<td>12</td>
</tr>
<tr>
<td>System use tutorial</td>
<td>Trainees watch a video demonstration of the EHR functions, and have hands-on practice of each function.</td>
<td>15</td>
</tr>
<tr>
<td>EHR and clinical process integration tutorial</td>
<td>Trainees watch a video demonstrating how to integrate the EHR functions into the clinical process, have hands-on practice, and get feedback</td>
<td>13</td>
</tr>
<tr>
<td>EHR Case 2</td>
<td>Trainees complete a more complicated patient case using the EHR</td>
<td>12</td>
</tr>
</tbody>
</table>

**Implementation and Evaluation**

We implemented the final version of training program in the nursing school Class Lab for 9 groups of trainees (n = 18 in total); each group had 1 to 3 trainees based on their availability. After training, we evaluated key aspects of the program by analyzing their performance data, and conducted interviews to gather their opinions about the training. Every session progressed as expected. Most trainees managed to complete the training slightly faster than scheduled.

Preliminary analysis of their performance showed the training was effective as measured by trainees’ pre-post training performance comparison: 50% of trainees were able to correctly document the intended medication in case 1 (pre-training) compared with 100% of trainees able to correctly document the intended medications in the more complex case 2 (evaluation). Trainees were able to do it much faster too: for those able to complete case 1, it took them on average 532 seconds (95% CI = 455-609); Trainees took on average 430 seconds (95% CI = 373-487) to complete the more complex case 2. The reduced variance in duration is remarkable, because it is considered as one important indicator of an effective training, even with more trainees completing the task (n = 9 vs. n = 18). The results of their performance are summarized in Table 3.

**Table 3.** Performance metrics

<table>
<thead>
<tr>
<th></th>
<th>Pre Training Session (Case 1)</th>
<th>After Training Session (Case 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Trainees who completed the case</td>
<td>9 out of 18</td>
<td>18 out of 18</td>
</tr>
<tr>
<td>Duration of completion (in seconds with 95% confidence interval)</td>
<td>532 (77)</td>
<td>430 (57)</td>
</tr>
</tbody>
</table>

In addition to the quantitative measures, we conducted interview with trainees to gather their opinions about the program. We asked them to comment in terms of usefulness, clarity and any other aspects related to the training quality, and we recorded their comments. Their feedback was highly positive, and some frequent quotes include:

- “It (the training process) is smooth and well-planned.”
- “They (the materials) are really clear and well explained.”
Discussion

Instructional models, such as ADDIE or R2D2, can be helpful to guide a successful EHR training program development process. They help identify and remedy potential issues before the training program execution, to avoid potential costly failures and do-overs. It is a best practice to involve targeted trainees in the process, and understand their work, needs, characteristics, and current competencies, and update the training materials and processes to best serve their needs. However, the training design and development process can be frustrating during early stages; for example, during our early development phase, a trainee involved in testing told us, “do not teach us how to use the (cumbersome) system, go fix it.” While an ideal and user friendly system is the ultimate goal, a good training program is necessary for users to take full advantage of any EHR system.

In order to make progress toward the correct direction during design and development process, different models all suggest conducting extensive analysis to map out the specifics of the planned training, such as trainees, resources, training goals. Based on the analysis results, we can go through iterations of prototyping, testing, redesign and development until we are confident the program is ready to be implemented. It is difficult to have a perfect training program through just one round of development, and we can always learn new things based on different feedback. The trainees who involved in the development process could also serve as super user instructors to other trainees.

The development of an EHR training program should focus on content and process, which are both key factors to ensure success. The content and process are interactive factors, as different content may be best trained using different methods or durations. Our experience was to emphasize more on content clarity and learnability during early stage, and focus more on process toward later stage to avoid trainees’ loss of attention and fatigue, such as self-pace learning, video demonstration vs. paper based content presentation. The model presented in this paper provides guidelines and principles for the design of content and process, and can be applied to various settings and tasks. However, different training programs may vary in detailed content and process based on training needs and resources, and everything should be vigorously tested before implementation.

References


Accelerating Chart Review Using Automated Methods on Electronic Health Record Data for Postoperative Complications

Zhen Hu, ME1, Genevieve B. Melton, MD, PhD1,2, Nathan D. Moeller, MS3, Elliot G. Arsoniadis, MD1,2, Yan Wang, PhD1, Mary R. Kwaan, MD, MPH2, Eric H. Jensen, MD2, Gyorgy J. Simon, PhD1,4
1Institute for Health Informatics, 2Department of Surgery, 3Department of Computer Science and Engineering, 4Department of Medicine, University of Minnesota, MN

Abstract

Manual Chart Review (MCR) is an important but labor-intensive task for clinical research and quality improvement. In this study, aiming to accelerate the process of extracting postoperative outcomes from medical charts, we developed an automated postoperative complications detection application by using structured electronic health record (EHR) data. We applied several machine learning methods to the detection of commonly occurring complications, including three subtypes of surgical site infection, pneumonia, urinary tract infection, sepsis, and septic shock. Particularly, we applied one single-task and five multi-task learning methods and compared their detection performance. The models demonstrated high detection performance, which ensures the feasibility of accelerating MCR. Specifically, one of the multi-task learning methods, propensity weighted observations (PWO) demonstrated the highest detection performance, with single-task learning being a close second.

Introduction

Conducting research and quality improvement using manual chart review (MCR) remains widely used in traditional observational clinical studies aimed at assessing detailed information on patients to understand disease course or outcomes and is also a primary modality used for quality improvement, epidemiologic assessments, and for graduate and ongoing professional education and assessment.1-5. Prominent examples of healthcare quality improvement programs include the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) and Centers for Disease Control and Prevention (CDC)’s National Healthcare Safety Network (NHSN), which employ MCR to retrospectively measure healthcare quality and patient safety outcomes.6 For example, NSQIP surgical clinical reviewers perform MCR to collect and report 21 surgical adverse events and related preoperative, intra-operative, and postoperative clinical data elements. Similarly, NHSN is a required healthcare-associated infection (HAI) tracking system, which also relies significantly on MCR. The reviewers for both of these programs utilize data from the electronic health record (EHR) and any paper records, including results of diagnostic tests, diagnoses information, and narrative text to ascertain outcomes.6-8 Though MCR provides high-quality data for further secondary purposes including research, it is very time-consuming and labor-consuming to conduct and may be a bottle-neck step in the research discovery process.

Among all postoperative complications, the most common types are surgical site infection (SSI), pneumonia, urinary tract infection (UTI), and sepsis, accounting for nearly 60% of all complications.9 Severe infections could trigger sepsis and even septic shock, particularly in people who are already at risk. Sepsis and septic shock are common and deadly, and CDC has listed “septicemia” as the 11th leading cause of death nationwide.8 In addition, postoperative complications are expensive to treat. According to a recent study, 440,000 of these adverse events happen annually and cost overall up to 10 billion dollars per year in the United States.9 Given the significant influence on the quality and cost of healthcare, postoperative complications are increasingly and widely viewed as a quality benchmark and are a strong emphasis of national initiatives for infection prevention and control.10 However, infection control departments at medical facilities spend considerable time and resources on MCR to collect infection outcome data for public reporting and surveillance, which has greatly increased the burden on already limited infection prevention and quality improvement resources.

To improve the efficiency of MCR, some groups have explored the feasibility of computerizing the process of MCR for reporting postoperative complications or for automatically collecting necessary data elements in quality improvement program, but these efforts have mainly been based on unstructured data such as physician narratives and nursing notes.11-14 In one example, relevant keywords and phrases were extracted from free text documents for adverse event detection.13 In another case, statistical and rule-based extractors were developed to automatically
abstract data elements such as procedure type and demographic information from clinical notes\textsuperscript{14}. According to standard NHSN definition, SSIs can be categorized into superficial, deep, and organ space\textsuperscript{15}. In our previous work, three subtypes of SSI and the overall SSI detection models were developed based on structured EHR data, including lab tests, vital signs, medications, and orders\textsuperscript{16}. These models have high specificity as well as very high negative predictive values, guaranteeing the vast majority of non-SSIs could be eliminated thus significantly reducing the burden on chart reviewers.

In this study, our aim is to build an automated platform for postoperative complications detection based on structured EHR data by using robust modeling techniques. Included in this analysis are the main postoperative complications of three subtypes of SSI (superficial, deep, and organ space), pneumonia, UTI, sepsis, and septic shock. We hypothesized that EHR data would include significant indicators and signals of postoperative complications and that sophisticated machine learning methods might be able to extract these signals, accelerating the MCR process of these adverse events. Compared with the gold standard MCR process, automated application has potential advantages. First, MCR lacks inter-rater reliability, while an automated abstraction system would provide an objective and consistent reporting protocol that can be applied across multiple medical institutions. Second, a successful automated abstraction system would allow for expansion to include other procedures where postoperative surveillance is not being performed currently.

Specifically, we explore several methods for developing postoperative complication detection models. The most straightforward way to detect each type of complications is to build an independent classifier for each of them, which could be viewed as single-task learning, where detecting each complication is a \textit{task}. When tasks are known to share similar features, we expect the resultant models to be similar. Learning models for these tasks together allows us to introduce inductive bias to make the resultant models similar, providing us with more robust models. Learning models for related tasks together is referred to as multi-task learning. For example, when our task is to identify a particular SSI subtype, a related task can be to detect any SSI (overall SSI). With an overall SSI model in hand, identifying a particular SSI subtype is easier, because the classifier only needs to learn the difference between overall SSI and the particular SSI subtype, rather than the difference between the SSI subtype and any other complication.

In our application, we have a hierarchy of tasks as shown in Figure 1. The first task is to distinguish patients with infection from those without. Next we distinguish among the various kinds of infections and finally, if the patient happens to have SSI, we distinguish among the three types of SSI. We assume that many infections share some characteristics that other diseases do not; and we further assume that many types of SSI share some characteristics that non-SSI infections do not. Our hypothesis is that by making a task-similarity hierarchy available to the multi-task learning methods as domain knowledge, they can utilize these information towards building more robust and better performing detection models.

![Figure 1. A hierarchical structure among postoperative complications](image)

In this work, we compare six methods for developing post-operative complications detection models. First we have single task learning, where predicting each complication is an independent task. We also explore five different methods for multi-task learning, assessing their value in improving the detection performance.
Materials and Methods

The overall methodological approach for this study included four steps: (1) identification of surgical patients and collection of associated data, (2) data preprocessing, (3) supervised single-task and multi-task learning model development, and (4) evaluation of a series of final models using gold standard data from the NSQIP registry. Institutional review board approval (IRB) was obtained and informed consent waived for this minimal risk study.

Data collection

Surgical patients at the University of Minnesota Medical Center are annually collected following strict inclusion and exclusion criteria. Their occurrences of postoperative complications are extracted and documented by chart reviewers, which are used as the gold standard in this study. We first identified surgical patients from 2011 to 2014 who had been selected for inclusion into the ACS-NSQIP. Clinical data for the identified patients were extracted from our clinical data repository (CDR) and their postoperative complication outcomes were retrieved from the NSQIP registry. The dataset was divided into training set (first 2.5 years) for model development and test set (last 1.5 years) for evaluation. The occurrences of postoperative complications in both training and test set are shown in Table 1. Overall SSI included any type of SSI.

We collected demographic information (e.g., age, gender), laboratory results (e.g., white blood cell count, glucose), microbiological results (e.g., urine culture, blood culture), relevant diagnosis codes (e.g., ICD_9 code 599.0 for UTI, ICD_9 code 995.91 for sepsis), orders and procedures for diagnosis and treatment (e.g., chest radiological exams for making diagnosis of pneumonia, CT guided drainage to treat SSI), antibiotic use, vital signs (e.g., temperature, heart rate), and medications.

Table 1. Postoperative complications distribution in training and test set

<table>
<thead>
<tr>
<th></th>
<th>All Complications</th>
<th>Superficial SSI</th>
<th>Deep SSI</th>
<th>Organ Space SSI</th>
<th>Overall SSI</th>
<th>Pneumonia</th>
<th>UTI</th>
<th>Sepsis</th>
<th>Septic Shock</th>
<th>All Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>571</td>
<td>168</td>
<td>76</td>
<td>105</td>
<td>336</td>
<td>124</td>
<td>140</td>
<td>115</td>
<td>34</td>
<td>5280</td>
</tr>
<tr>
<td>Test set</td>
<td>279</td>
<td>59</td>
<td>26</td>
<td>73</td>
<td>157</td>
<td>48</td>
<td>76</td>
<td>30</td>
<td>17</td>
<td>3629</td>
</tr>
</tbody>
</table>

Data preprocessing

Data preprocessing generally included data cleaning and missing data imputation. Our previous work on missing data imputation methods suggests that filling in the average value of patients without complications (normal value) in the training set and the test set introduces the least bias\(^7\). Accordingly, in this research, we chose to follow this imputation method.

For longitudinal lab results and vital records, we used aggregated features. We included the most recent value before the operation as the baseline, and the extreme and mean values from day 3 to day 30 after the operation during follow-up. We assembled a list of relevant antibiotics for each specific outcome, and extracted different classes of antibiotics as features. For relevant orders, procedures, and diagnosis codes, binary variables were created to denote if they were assigned to a patient. Additionally, we not only considered whether a microbiology test was ordered or not, but also looked at the specific bacterial morphological types (e.g. gram positive rods, gram negative cocci). Two binary features were built for each test to represent a culture placed or not and a positive/negative result, separately.

Modeling Methods

After data collection and preprocessing, the six modeling techniques were applied.

When building our detection models, we face three key challenges. The first one is the skewed class distribution. A mere 10% of patients in the training set have any complications and some complications, like septic shock, occur in only .6% (half a percent) of the patients. The second challenge is the small sample size. Some complications, like septic shock, only have 34 observations in the training and 17 in the test set. While our problem does not appear particularly high dimensional, for these rare complications, the number of predictors (approx. 200) exceeds the number of samples. The third challenge is the heterogeneity of the outcomes. We have 9 outcomes, each having their own specific characteristics and there are also variations among patients who do not have any complications. A successful detection algorithm has to address some of these challenges.

Below, we explain each of the six methods and describe which of the above challenges they address.
**Method 1: Single-task learning**

The nine outcomes are modeled independently using Lasso-penalized logistic regression. Lasso-penalized regression constructs a sparse model, where some of the coefficients are set to exactly 0. This performs automatic variable selection and also helps with the small sample size for some of the outcomes. This method does not specifically address the other challenges.

**Method 2: Hierarchical classification**

Let us consider the hierarchical structure among surgical patients as shown in Figure 1. All tasks are divided into three levels and models are constructed in a top-down fashion. The top-most task identifies patients with any postoperative complication. Next, in patients who are predicted to have complications, a 2nd level task is carried out to distinguish between SSI, pneumonia, UTI, sepsis, and septic shock. If a patient is predicted to have SSI, a further 3rd level task is also carried out to identify the SSI type: superficial, deep, and organ space SSI. Each task utilizes Lasso-penalized logistic regression. When more than two classes are possible, the one-vs-all approach is used to break a multi-class classification into a set of binary classifications.

As the method progresses from the top towards the bottom of the hierarchy, it gradually focuses on subpopulations that are enriched in the outcome of interest. This addresses heterogeneity by explicitly ignoring patients without indication of the outcome and also addresses the skewed class distribution. The adoption of LASSO model can overcome the problem of small sample size.

**Method 3: Offset method**

Similar to the hierarchical method, the classifiers for different tasks are built in a top-down fashion. For the top level task (i.e. complication classifier), a LASSO logistic regression classifier is built directly. For the lower-level tasks, we essentially model the difference between the parent and the child task. For example, the deep SSI classifier models the difference between overall SSI and deep SSI. This is achieved through penalizing the child model against the parent model: the predictions from the parent model are included as an offset term (a term with fixed coefficient of 1) in the child model, which is a LASSO logistic regression classifier. Due to the Lasso penalty, variables that have the same effect in the parent and child model will have a coefficient of 0; and conversely, variables that have non-zero coefficient are the variables in which the parent and child tasks differ. The method addresses the challenge of small sample size in two ways. First, in contrast to method 2, it uses the entire population at each level, thus the problem does not become overly high dimensional. Also the offset biases the child classifier towards the parent model. This method only offers limited ability to address heterogeneity.

**Method 4: Propensity weighted observations (PWO)**

The propensity weighted observations method also builds classifiers from the top level to the bottom level. The classifier of the top-level task, the complication classifier, is LASSO-penalized logistic regression (same as all of the previous methods). The classifiers for the second level task are built on the entire population, however, the observations (patients) are weighted by their propensity of having a complication. The propensity is obtained from the higher-level (complication) classifier. Patients, who are likely to have a complication receive a relatively large weight, while patients who are unlikely to have a complication receive a small weight. Therefore, patients with complication contribute more to the 2nd level classifiers than those who are unlikely to have complications. Similarly, the 3rd level classifiers, which distinguish between the three kinds of SSI, are also built on the entire population. The weights of the patients are their propensity of having SSI, thus the patients who likely have SSI contribute more to these classifiers than patients who are unlikely to have SSI. Similarly to the offset method, the PWO method uses the entire population, but by applying weights, it reduces outcome heterogeneity (patients with unrelated complications receive small weights) and reduces the skew of the class distribution by enriching the training set with patients having the outcome of interest (these patients receive high weights).

**Method 5: Multi-task learning with penalties (MTLP)**

Unlike the previous methods, the objective of multi-task learning with penalty (MTLP) method is to learn the regression coefficients $\beta_i$ for all tasks simultaneously. In the MTLP method, we assume that the parent task and its child tasks share some features and the respective models should have similar coefficients for those features. Similarly, to the offset method, this similarity is enforced through penalizing the child model against the parent model. Unlike the offset method, which builds models in a top-down manner, MTLP builds all models simultaneously. Specifically, the objective function is
argmin_{\beta \in \mathbb{R}^p} l(\beta_t) + r_1(\beta_{level,2}) + r_2(\beta_{level,3})

(1)

It consists of three parts, the negative log likelihood of logistic regression, \( l(\beta_t) \), and two regularization terms, \( r_1(\beta_t) \) and \( r_2(\beta_t) \), as shown below.

\[
l(\beta_t) = -\left[ \frac{1}{T \times N_t} \sum_{i=1}^{T} \sum_{t=1}^{N_t} \left( y_{t,i} \cdot (x_{t,i}^T \beta_t) - \log(1 + e^{x_{t,i}^T \beta_t}) \right) \right]
\]

(2)

\[
r_1(\beta_t) = \lambda_1 \sum \| \beta_{level1 \ parent \ task} - \beta_{level2 \ children \ tasks} \|
\]

(3)

\[
r_2(\beta_t) = \lambda_2 \sum \| \beta_{level2 \ parent \ task} - \beta_{level3 \ children \ tasks} \|
\]

(4)

where \( T \) and \( N_t \) are the number of tasks and training set for each task, respectively; \( x_{t,i} \) and \( y_{t,i} \) are the feature vector and the label for the subject \( i \) in task \( t \), respectively; \( \beta_t \) is the coefficient vector for the task \( t \). The two regularization terms, \( r_1(\beta_t) \) and \( r_2(\beta_t) \), restrict the difference in coefficients between the level 1 parent task and its level 2 child tasks; and the difference between the level 2 parent task and its level 3 child tasks, respectively. Penalizing the difference between the parent and child models makes them similar. The MTLP method addresses heterogeneity by explicitly making the parent and child models similar, thereby essentially only modeling the difference between them; and it addresses the small sample size through the use of the entire population and regularization.

**Method 6: Partial least squares regression (PLS)**

As with the MTLP method, partial least squares (PLS) regression models all tasks simultaneously. PLS regression is similar to principal components regression in the sense that both methods reduce the dimension of input data by projecting the outcomes and predictors into new spaces and then build regression models in those new spaces. PLS differs from MTLP in that the task hierarchy is not explicitly given to the fitting algorithm; the algorithm has to autonomously learn the relationships among the tasks.

**Evaluation**

Outcomes based on MCR from ACS-NSQIP were used as gold standard to be compared with the results of postoperative complication detection models. The evaluation metric is area under the curve (AUC), which is commonly used to compare detection models. The range for AUC is between .5 and 1, .5 indicating a random model and 1 indicating perfect discrimination among the outcomes. We report the cross-validated AUC on the training set. To assess the variability of the detection performances on the test, bootstrap replication was applied and the 95% (empirical) confidential interval (CI) and mean AUC scores are reported, as well. Since all methods were evaluated on the same bootstrap samples, paired t-test was used to compare each pair of methods and assess the statistical significance of the observed differences in performance.

**Results**

**Evaluation results of six detection methods**

Figure 2 depicts the performances of the six methods. Each plot in Figure 2 corresponds to a task (complication) and each column in each plot corresponds to a method. Methods are numbered in the same order as they appear in the Methods section: #1 corresponds to Single-task, #2 to Hierarchical, #3 to Offset, #4 to Propensity Weighted Observations (PWO), #5 to Multi-Task Learning with Penalty (MTLP), and #6 corresponds to Partial Least Squares (PLS). The vertical axis is AUC. For each method, the mean AUC (across the bootstrapped test samples) is represented by a disk and lines extending out of the disk correspond to the 95% CI.

To assess the statistical difference between some of the methods, in Table 2, we show the results of pairwise (paired) t-tests among the various methods. The rows of the table correspond to tasks, the columns to a comparison between two methods. Each cell contains a number, which indicates which method has a significantly better performance and we also provide the p-value in brackets. ‘NS’ means ‘not significant’. The methods are numbered in the same way as above.

For the detection of all complications, Single-task, Hierarchical, Offset, PWO, and MTLP have the same good performance, and are significantly better than PLS. To detect superficial SSI, Offset and PWO have virtually identical performance (difference is not significant) and they perform significantly better than the other four methods. PWO performs best for detecting deep SSI and overall SSI. Single-task, PWO, and MTLP all perform
similarly (no statistically significant difference) in detecting organ space SSI but perform significantly better than the other methods. To detect pneumonia and UTI, Single-task and MTLP are not significantly different from each other but are significantly better than other four methods. Single-task, PWO, and MTLP are the top three in detecting sepsis and PWO is also the best method for detecting septic shock. In general, PWO is the best method for detecting most complications. Single-task and MTLP are close seconds (and they have virtually identical performance) and PLS is the method with the worst overall performance.

Figure 2. Detection performance of six models for all nine tasks, showing the mean and 95% CI
Detailed information about the performance of the methods is depicted in Figure 2 and the statistical significance of the pairwise comparisons between the various methods is shown in Table 2.

Table 2. Paired t-test results to compare different methods

<table>
<thead>
<tr>
<th>Category</th>
<th>Method 1 vs. 2</th>
<th>Method 1 vs. 3</th>
<th>Method 1 vs. 4</th>
<th>Method 1 vs. 5</th>
<th>Method 2 vs. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superficial SSI</td>
<td>2 (&lt;2.2e-16)</td>
<td>3 (&lt;2.2e-16)</td>
<td>4 (&lt;2.2e-16)</td>
<td>NS</td>
<td>3 (&lt;1.533e-14)</td>
</tr>
<tr>
<td>Deep SSI</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (=5.073e-07)</td>
<td>4 (&lt;9.285e-10)</td>
<td>NS</td>
<td>3 (&lt;2.2e-16)</td>
</tr>
<tr>
<td>Organ Space SSI</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (&lt;2.2e-16)</td>
<td>NS</td>
<td>NS</td>
<td>3 (&lt;2.2e-16)</td>
</tr>
<tr>
<td>Overall SSI</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (&lt;2.011e-13)</td>
<td>4 (&lt;3.572e-15)</td>
<td>NS</td>
<td>3 (&lt;2.2e-16)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (=2.353e-8)</td>
<td>NS</td>
<td>2 (&lt;2.2e-16)</td>
</tr>
<tr>
<td>UTI</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (&lt;2.2e-16)</td>
<td>NS</td>
<td>2 (&lt;2.2e-16)</td>
</tr>
<tr>
<td>Sepsis</td>
<td>1 (&lt;2.2e-16)</td>
<td>1 (&lt;2.2e-16)</td>
<td>NS</td>
<td>NS</td>
<td>2 (&lt;2.2e-16)</td>
</tr>
<tr>
<td>Septic Shock</td>
<td>NS</td>
<td>1 (&lt;2.2e-16)</td>
<td>4 (&lt;1.671e-12)</td>
<td>NS</td>
<td>2 (&lt;2.2e-16)</td>
</tr>
</tbody>
</table>

Table 3. Selected important variables for all complications and their descriptions

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis code</td>
<td>ICD_9 code: 998.59 and 997.32</td>
<td>Diagnosis code of postoperative SSI and pneumonia</td>
</tr>
<tr>
<td>Microbiology test order</td>
<td>Abscess culture</td>
<td>These are binary features to indicate if such microbiology test ordered or not during day 3 to day 30 after operation.</td>
</tr>
<tr>
<td></td>
<td>Blood culture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gram stain culture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sputum culture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Urine culture</td>
<td></td>
</tr>
<tr>
<td>Microbiology test result</td>
<td>Escherichia. Coli</td>
<td>These are binary features to indicate if the type of bacteria is positive or not matter in which kind of microbiology test during day 3 to day 30 after operation.</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus</td>
<td></td>
</tr>
<tr>
<td>Antibiotic use</td>
<td>Antibiotic_Superficial_SSI</td>
<td>They are binary features to indicate if antibiotics is placed to patients during day 3 to day 30 after operation.</td>
</tr>
<tr>
<td></td>
<td>Antibiotic_Pneumonia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Antibiotic_UTI</td>
<td></td>
</tr>
<tr>
<td>Laboratory results</td>
<td>Measurement_CR</td>
<td>The number of measurements for creatinine, platelet count test, prealbumin, and urine white blood cells (WBCU).</td>
</tr>
<tr>
<td></td>
<td>Measurement_PLT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Measurement_PREALAB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Measurement_WBCU</td>
<td></td>
</tr>
</tbody>
</table>

**Significant variables selected**

Lasso-penalized regression performs automatic feature selection. In Table 3, we provide a list of the most important features selected by the model that aims to identify whether a patient has a complication. This model is common across most methods. Due to space limitation, we cannot provide a list for all methods and all complications. Below we provide some examples of features selected by the best performing method for each of the complications.
**Superficial SSI** detection model based on Offset and Propensity Weighted Observations methods selected antibiotic use, gram stain ordered, and the ICD_9 code of SSI.

Besides diagnosis codes, antibiotics use, gram stain culture, **deep SSI** detection model based on PWO selected more features from laboratory results (mean value of creatinine and maximum value of WBC) and two more microbiology tests (tissue and wound culture).

**Organ space SSI** models based on Single-task, PWO, and MTLP methods have quite similar detection performance and important variables. The selected variables include four features of bacteria type (streptococcus, gram positive cocci, enterococcus, and escherichia. coli), three microbiology cultures (abscess and fluid culture), and the imaging orders of treatment. Interestingly, the diagnosis code of sepsis and imaging orders of sepsis treatment are selected as well. These can be explained by the fact that there are over 30 patients in our cohort have sepsis and organ space SSI together.

**Overall SSI** detection models based on weighted observation and offset methods perform with no significant difference. The important variables selected are antibiotic use (UTI, superficial and deep SSI), microbiology cultures (abscess, fluid, and wound culture, and the gram stain test), two types of bacteria (escherichia. coli and staphylococcus) and two relevant order features (imaging orders for diagnosis and procedures of treatment).

Besides the diagnosis code and antibiotic use, **pneumonia** models based on Single-task and MTLP include two features from microbiology test (bronchial and sputum culture), one binary feature of image-guided diagnosis orders, and two aggregate features from lab tests (the mean value of PCAL and PH).

**UTI** models based on Single-task and MTLP selected diagnosis codes (for UTI), antibiotic use, placement of urine culture, and two bacteria types (proteus and escherichia. coli).

For **sepsis** models, the top performing methods, Single-task, PWO, and MTLP, selected features including antibiotic use, microbiology cultures (abscess, blood, fluid, and urine culture, and the gram stain), two bacterial types (enterococcus and escherichia. coli), and the image guided orders of treatment.

For **septic shock** detection, most models only selected diagnosis codes. However, PWO included more variables, such as laboratory tests (maximum value of partial thromboplastic time, PH, mean value of lactate), bacteria types (stentotrophomnas and staphylococcus) and the tracheal culture.

**Discussion**

Manual chart review for post-operative complications is very resource intensive. In this work, we examined whether EHR-based state-of-the-art predictive modeling approaches can learn characteristics of various types of complications and subsequently detect them reliably. With detection performances (measured as AUC) exceeding 0.8 for all complications and even 0.9 for some complications, the answer is affirmative: machine learning detection models definitely have the potential to help detect post-operative complications automatically. The question is which modeling approach is best suited for this application.

Post-operative complications are heterogeneous; they cover a wide-range of conditions, each having their own diagnostic methods, diagnoses codes, laboratory tests, and diagnostic and therapeutic procedures. They can be organized into a hierarchy and complications on the same level of the hierarchy are more similar to each other than to complications on a higher level of the hierarchy. Multi-task learning methods have the ability to exploit such similarities towards achieving better detection performance and more stable models even when the sample sizes are small.

We compared six approaches to building post-complication detection models. One of them was single-task learning, where we build independent models for each task; four methods were multi-task learning methods that can utilize the hierarchy of complications; and finally, we also utilized Partial Least Squares (PLS), which can simultaneously model multiple outcomes, but it tries to autonomously detect the relationship among the outcomes. PLS thus stands in sharp contrast with the other multi-task learning methods, as PLS automatically infers the relationships among the complications, while the other multi-task learning methods receive this information from an expert.

We found PLS to have the overall worst performance. This is not surprising, since PLS receives less information than the other multi-task learning methods. We expect PLS to bias the models based on the relationships among the outcomes, but we do provide it with these relationships. If PLS infers the relationships among outcomes incorrectly, it will bias the models incorrectly, eroding detection performance. With some of the complications having small sample sizes, it is unsurprising the PLS failed to infer the correct relationships. If we had substantially more samples,
Hierarchical modeling also had disappointing performance. The essence of hierarchical modeling is to build classifiers in a subpopulation that is greatly enriched in the outcome of interest. For example, distinguishing among the three types of SSI is easier in a subpopulation of SSI patients than it is in the general population. The performance of the method did not live up to our expectation for two reasons. First, while these outcomes are rare (deep SSI occurred in 76 patients out of 5280), they still occur in sufficient numbers for a Lasso-penalized logistic regression model. The second reason concerns the way the subpopulations were constructed. If the higher-level classifier (does this patient has SSI?) predicts the patient to be free of SSI, then this patient does not enter the subpopulation and the deep SSI detector has no opportunity to learn from this sample. We could have built the deep classifier on the true SSI patients (rather than the predicted SSI patients), but then the distributions of the training SSI patients (true SSI patients) and the test SSI patients (predicted SSI patients) would be different, leading to degraded detection performance. Our results with the Propensity Weighted Observations method tell us that the concept of enriching patients with SSI for (say) the deep SSI classifier is valid; the hierarchical method simply implemented this concept suboptimally.

The Propensity Weighted Observations (PWO) method achieved the overall highest performance with a margin that is statistically significant. PWO is closely related to the hierarchical method in that it enriches the training sample with patients who have the outcome of interest. In contrast to the hierarchical method, it achieves this enrichment through constructing a new sample, which is a propensity weighted version of the original population. For example, to identify patients with deep SSI, PWO uses the entire population, but patients with high propensity for SSI receive high weight and patients with low propensity for SSI receive low weight. The hierarchical method is a binary version of PWO, where the weights are either 0 or 1. Having the propensity weighted population removes the problem of excluding patients based on an incorrect prediction. Suppose our SSI classifier misclassifies a deep SSI patient as not having SSI. This patient will still be included in the training set for the deep SSI classifier; this observation will receive a slightly lower weight. Admittedly, using propensity score weighing for multi-task learning is rather unusual; we did not expect this method to perform so well.

The offset method, like PWO, always uses the entire population for classification. Its performance falls short of that of PWO, because its ability to remove heterogeneity is limited. It can bias a child model against the parent model, which helps with small sample sizes (it performed well on superficial SSI), but has limited effect on removing the variation in (say) normal patients. PWO is more effective at removing heterogeneity: patients with unrelated complications receive a low weight and contribute to the model only minimally.

MLTP is essentially identical to the offset method, except MLTP optimizes all outcomes simultaneously (as opposed to sequentially in a top-down manner). In a top-down construction scheme, only the parent task can influence the child task; the model from the child task cannot influence the parent model. When all tasks are carried out simultaneously, the child models can influence the parents, as well. As a result, MLTP was either the best or second best method for almost all rare (<3% of patients) outcomes. The caveat of simultaneous optimization is the increased potential for overfitting. Indeed, comparing MTLP’s cross-validated AUC scores on training set to those on the test set, reveal signs of overfitting. For example, to detect sepsis, it has a very high training AUC (>0.96), but the 95% CI of AUC on test set is only (0.8986, 0.9183).

Conclusion

Developing machine learned models to automatically detect post-operative complications definitely has the potential to accelerate the manual chart review process. We found that multi-task learning, specifically, the propensity weighted observations method, statistically significantly outperformed the single-task learning approach. While the difference in detection performance was relatively modest (albeit significant), the additional cost of implementing this method over the standard single-task learning method is minimal. Thus, we would recommend trying both single-task learning and PWO.

Our application was relatively easy: we had sufficiently many samples for Lasso-penalized logistic regression to construct a good model even for the most infrequent outcome. In an application, where fewer samples are available or outcome distributions are more skewed, we would expect the performance gap between multi-task learning and
single-task learning to open up, providing a more attractive implementation cost versus detection performance proposition for multi-task learning.

Our future work includes building postoperative complications detection models using both structured and unstructured EHR data. We hypothesize that the combination of structured and unstructured clinical data would include more significant indicators and signals of postoperative complications, and improve the performance of detection. The performance of the models with only structured data and that of the models with both structured and unstructured data will be compared and evaluated.

Acknowledgements

The National Institutes of Health through the National Library of Medicine (R01LM011364 and LM011972-01A1), Clinical and Translational Science Award (8UL1TR000114-02), and University of Minnesota Academic Health Center-Faculty Development Grant supported this work. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Reference

Development of a Novel Markov Chain Model for the Prediction of Head and Neck Squamous Cell Carcinoma Dissemination

Hyunggu Jung, MS, MMATH\(^1\), Anthony Law, MD, PhD\(^2\), Eli Grunblatt\(^3\), Lucy L. Wang, MS\(^1\), Aaron Kusano, MD, MS\(^4\), Jose L. V. Mejino Jr., MD\(^5\), Mark E. Whipple, MD, MS\(^1,2\)

\(^1\)Department of Biomedical Informatics and Medical Education
\(^2\)Department of Otolaryngology-Head and Neck Surgery
\(^3\)Medical Scientist Training Program
\(^4\)Department of Radiation Oncology
\(^5\)Department of Biological Structure
University of Washington, Seattle, WA

Abstract

Prediction of microscopic tumor spread to regional lymph nodes can assist in radiation planning for cancer treatment. However, it is still challenging to predict tumor spread. In this paper, we present a unique approach to modeling how tumor cells disseminate to form regional metastases. This involves leveraging well established knowledge resources and commonly held notions of how cancer spreads. Using patient data, we utilized our approach to create a model of metastasis for the subset of head and neck squamous cell carcinoma that arises in the mucosa of the lateral tongue. The model was created using a training set extracted from the clinical records of 50 patients with tumors of this type who presented to the University of Washington head and tumor board over a three and half year period. The test sets consist of four case series drawn from the literature.

Introduction

Head and neck squamous cell carcinoma (HNSCC) is the most common form of cancer of the upper aerodigestive tract. The mainstays of treatment consist of surgery and/or radiation therapy. Radiation therapy involves identifying the target region in the patient (the neoplastic area requiring treatment) and then customizing a treatment plan that achieves a therapeutic dose of radiation to this zone while minimizing the dose to the surrounding area. Reports from the International Commission on Radiation Units and Measurements (ICRU)\(^1,2\) define several kinds of volumes to guide radiotherapy planning. The Gross Tumor Volume (GTV) is defined as “the gross palpable or visible/demonstrable extent and location of malignant growth.” The GTV is typically defined using computer drawing tools on computed tomography (CT) or other imaging modalities by an attending physician. Even with the best imaging technique, the boundaries of such a volume may be very uncertain, and much clinical judgement is involved. The Clinical Target Volume (CTV) is defined as the volume that includes the GTV and/or microscopic malignant disease. At present, the radiation oncologist uses his/her knowledge of the pathology literature, clinical experience and knowledge of anatomy to draw the CTV. This leaves much room for improvement, in particular in the ability to predict the extent of subclinical microscopic disease.

Prediction of microscopic spread of tumor cells is becoming critically important in the decision making process of planning radiation therapy for cancer. Until relatively recently, standard radiation treatment of HNSCC incorporated relatively large treatment areas. With the onset of Intensity Modulated Radiation Therapy (IMRT), treatment can be more precisely applied, such that if it is known that regional spread is confined to a specific set of lymph nodes, a more focused treatment can be considered. The resultant payoff is reducing or eliminating the collateral damage due to over-irradiation of surrounding tissue. Imaging methods such as CT or magnetic resonance imaging (MRI) can only show evidence of gross disease that has caused enlargement of the lymph nodes. Functional scans, such as positron emission tomography (PET), can demonstrate smaller metastases, but even with PET imaging it is not possible to determine with certainty whether nodes are involved and should be treated.\(^7\) None of these imaging modalities can detect micro-metastases, which can currently only be confirmed post-surgery by pathologic evaluations of surgical specimens.

The initial work on this project\(^8,9\) showed that a model of tumor dissemination based on an anatomy ontology and some simple assumptions can produce predictions that have a fair agreement with published surgical data. The modeling strategy uses the idea of a Markov chain. Each lymphatic chain\(^1\) can be described in terms of the probability

\(^1\)We are using the term “chain” in two ways here. A lymphatic chain is so called because it typically consists of a series of vessels connecting several lymph nodes in a linear string. That unit as a whole is modeled as a Markov chain, meaning its state is represented as a stochastic process, or sequence over time.
for it to be in any of five possible states, numbered 0 through 4. State 0 means no tumors present. State 1 means a little tumor, state 2 somewhat more, and so on, to state 4 which is a big palpable node. The only way a lymphatic chain can go from state 0 to state 1 is by some tumor cells migrating downstream from the upstream chain, but once some tumor cells are there, they can transition to the states of larger growth on their own. The simplest way to think about the spread of tumor cells is to imagine that the process happens in discrete time steps. The tumor is in some initial state, and each lymphatic chain is as well. The states change with each time step.

The idea of a Markov chain is that a system can undergo state transitions. Lymphatic chains can go from state 1 to state 2, and so on, with some transition probability, and the probability that the system is in any state at time \( t + 1 \) depends only on the present state, the state of everything at time \( t \), and not the history of how it got there.

If each lymphatic chain were independent of all the others, the probabilities would form a state vector with five components, one probability for each possible state of the lymphatic chain could be in. For each possible transition we would have some probability, \( p_{ij} \), the conditional probability that the system is in state \( j \) at time \( t + 1 \) given that it is in state \( i \) at time \( t \). Strictly speaking, this transition probability can depend on time. Processes in which the transition probabilities do not depend on time are homogeneous. In such cases, the transition probabilities form a matrix, and the relation between the state vector at time \( t + 1 \) and the state vector at time \( t \) is just multiplication of the state vector by the transition matrix, to get the new state vector.

In this paper, we introduce a new Markov chain model for nodal spread of head and neck cancer using data by T stage, tumor site and nodal stage to flexibly compare the output with patient data from surgical pathology of head and neck cancer. We decided to use the lateral tongue as our primary site of interest because it was a primary tumor for which we had a relatively large group of patients (50 patients) with which to find the parameters of the model, and were able to identify a number of studies in the literature that contained data by both T stage and lymph node involvement. To evaluate the accuracy of the model derived from the patient data, we systematically compare the predictions of the model for against the results found in the literature data.

**Methods**

**Assumptions**

We make a number of simplifying assumptions in building the current model. We assume that lymphatic chains are not “skipped” so that if disease is present in the fourth chain in the lymphatic path from the primary site, it is also present in the first three. Although skips do occur, they are relatively infrequent.\(^{10} \) We assume that the process of primary growth and the process of dissemination from one lymphatic chain to the next are independent of each other and their rates are independent of time. Since the density of lymphatic channels varies greatly by primary sites, we assume that the probability of initial metastasis may be different for different first echelon lymphatic chains. However, as an initial simplifying assumption, once a tumor is within the lymphatic system, we assume that the probability of dissemination from one chain to the next along a lymphatic path is the same for all paths and all chains. We also make a simplifying assumption that the probability of growth from one T stage to the next is the same for all T stage transitions. We assume that the probability of diagnosing a tumor may change as the size and metastatic spread of the tumor increases. We start with these simplifying assumptions, and find out from comparing model predictions with data how to refine this set of assumptions as the model is further developed.

**Structure**

Instead of modeling each chain as having several states with varying amounts of tumor, and using the T-stage as a proxy for how many iterations to run in the initial model,\(^8 \) we propose a richer but simpler model: a single Markov chain that represents the patient’s disease status using two layers of two dimensional array of states, indexed by T-stage for the primary tumor location and by the number of steps of metastasis along a lymphatic pathway as illustrated in Figure 1. The T-stage can take on 4 values, 1-4, corresponding to the standard clinical definition of T-stage for HNSCC. The metastatic progression along the lymphatics can be 0 for no metastasis, 1 for metastatic disease present in the first echelon of lymphatic vessels, 2 for disease present in the first and second echelons, and so on. For example, a patient with T-stage 2 and metastatic progression 2 has a primary tumor clinically defined as stage 2 and metastatic disease in the first two echelons of regional lymphatic nodes.

This describes a dynamic process whereby at each step a tumor can either grow in size to a larger T stage, metastasize further down the lymphatic pathway, or stay in the same state. With this model, given enough time, all tumors will invariably end up as the largest size with the most extensive metastasis. Of course, this is not what is actually observed in the clinical setting; at some point in this process the tumor is diagnosed, evaluated and a therapeutic intervention initiated. The size of tumor and degree of metastasis at the time of evaluation and/or management is what is typically
reported in the literature or clinical record. In order to model this process, we added an emission layer for each state, representing the probability of the tumor being diagnosed at that state. Once in the emission layer there is no further tumor growth or metastasis (the tumor has been diagnosed and recorded.) This approach provides several advantages. It allows us to model a clinical process that is inherent in the patient and literature datasets that we use to train and test the model.

The transition matrix for this set of states can be thought of as a four dimensional matrix indexed by the initial T-stage and lymphatic path depth, and the final T-stage and lymphatic path depth. Only stepwise transitions are hypothesized so the transition matrix is mostly zeros, and can be represented by a sum of delta functions. The definitions of parameters used in our Markov model are as follows: $g$ represents the probability of primary tumor growth by one T-stage. $m$ represents the probability of metastasis one step further down the lymphatic path. $h_1$ represents the probability of initial metastasis from primary site to one of the first echelon lymphatic chains, the Submandibular LC (Level I) in Figure 2. $h_2$ represents the probability of initial metastasis from primary site to the Superior cervical LC (Level II) as demonstrated in Figure 2, which is both a first echelon lymphatic chain via one pathway and a second echelon lymphatic chain from the Submandibular LC. $f$ represents a linear function with two variables, $\alpha$ and $\beta$, the probability of moving from a node to its corresponding emission layer node.

We represent the probability of being in state $s, c$ at time $t$ as $P_{s,c}(t)$ and the Markov chain then consists of the successive time states, where each state can be computed from the previous state by Equation 1.

$$P_{s,c}(t + 1) = \sum_{i=0, j=0}^{s, c} M_{s,c,i,j} P_{i,j}(t)$$

The sum is only taken up to the state indices of interest, since there is no contribution from a higher T-stage to a lower T-stage and there is no contribution from a downstream lymphatic chain to an upstream one.

**Training set and test set**

For our training set, we used data abstracted from the records of 50 patients who presented to the University of Washington head and neck tumor board over a three and half year time period with an initial diagnosis of non-treated
Figure 2: Diagram of lymphatic drainage pathways for the Lateral Tongue Tract. Arrows indicate sequential dissemination of metastatic tumor cells through lymphatic chains (LC). 12, 13, 14

...squamous cell carcinoma of the anterior tongue (excluding specific tumors of the base of tongue or ventral tongue.) This provided a dataset. Primary tumor stage was taken from the tumor board report and nodal involvement from the review of pathology reports of neck dissections. In some cases of low stage tumors (T₁ or T₂) without radiographic or clinical evidence of nodal involvement and very low suspicion of metastasis, a neck dissection was not performed. In these cases we assumed the absence of nodal metastasis. While the number of patients in the training set is smaller than some of the literature datasets, the patient dataset contains discrete information about the involvement of each nodal level at each T stage.

In the patient dataset that we used as a training set, the rate of metastasis was reported as the number of patients with a given T stage that had metastasis to a particular level of the neck, along with the number of total patients with that T stage. However, any individual patient may have metastasis to more than one level. In order to compare our model to the dataset, we converted our state probabilities to an analogous measurement. For each T stage, we determined the probability that a given echelon was involved with tumor by adding the probabilities of all states that included metastatic involvement of that echelon.

For our test sets, we selected four studies from the literature that record aggregated values for nodal involvement and T stage distribution. These studies were typically performed to test the utility of treatment modality, not necessarily to describe the patterns of lymph node metastasis. We chose two studies (Dogan, Dias) 5, 6 that provided aggregated information on all four potential lymph node levels. We also chose two studies (Lim, Sparano) 3, 4 which provided aggregated information on Levels I-III only. This was due to the fact that these studies utilized a type of neck dissection that does not include Level IV, due to a clinical decision that the probability of metastasis in these patients was too low to warrant removal of these lymph nodes. In these studies, we assumed that if Level IV was not treated then it did not contain metastases. (This correlates with our assumption in corresponding cases in the patient dataset.) While these literature studies do not provide discrete information about each individual nodal level by T stage, they have the advantage of larger numbers in some cases. Since each study draws patients from a somewhat different distribution of T stages, different studies allow us to test the predictive ability of our model across different T stage distributions.

Finding parameters
In order to execute the model, we need the parameters of m, g, h₁, h₂, α, and β values, in which m represents the probability of further metastasis along the lymphatic pathway, g represents the probability of primary tumor growth to a higher T stage, h₁ and h₂ represents the probability of initial metastasis to the two potential first echelon nodes, and α and β are coefficients in a linear function that represents the probability of diagnosing a tumor in that specific state. The strategy of finding parameters was to execute the model with a set of possible cases with a training dataset and get the optimal case where the output matrix is closest to the training dataset. We then identified the parameters based on the output of the case.

To generate a set of possible cases, we used a single layer grid search with a combination of m, g, h₁, h₂, α, and β values with the following ranges: m, g, h₁, and h₂, between 0 and 1 with a step size of 0.05 and α and β between 0 and 1 with a step size of 0.025. This generated 326,922, 561 (≈ 21 × 21 × 21 × 21 × 41 × 41) cases. We then excluded...
cases if the sum of the probabilities of the outgoing edges from each node was not 1 (one of the Markov chain rules).

After that, we judged closeness of the output of the case and training dataset by calculating the root mean square (RMS) error using the Equation 2. The distance between our model output and the training data is smaller when the RMS error is small. Let $A$ and $B$ be a matrix, $a_{ij}$ and $b_{ij}$ is an element with row $ai$ and column $j$. $C$ is the number of echelons and $S$ is the number of $T$-stages.

$$rms(A, B) = \sqrt{\sum_{i=1}^{C} \sum_{j=1}^{S} (b_{ij} - a_{ij})^2}$$

(2)

**Execution**

We ran the model with four literature data as test datasets using the parameters, $m$, $g$, $h_1$, $h_2$, $\alpha$, and $\beta$, we found from a training dataset. We ran our model until the sum of the probabilities on the emission layer becomes greater than 0.99. Prior to running the model, the initial probability matrix $P$ used in the Markov chain model for the initial tumor site is given below:

$$P_{init} = \begin{pmatrix}
1.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00
\end{pmatrix}$$

$P_{11}$ has probability 1, indicating the site of our primary tumor. All other probabilities are 0, indicating that there is no prior metastasis or tumor growth. The model is then run for $n$ iterations where $n$ is the number of the states. The number of iterations was the number of the states because our model calculates the final probability for each state. For each iteration, the probability being in each state is calculated using the transition matrix that corresponds to our two-dimensional model as illustrated in Figure 1.

**Results**

**Parameters**

After systematic runs with a range of parameter values, we identified 18 out of 188,039 (top 0.01 percent) combinations of parameters, $m$, $g$, $h_1$, $h_2$, $\alpha$, and $\beta$, based on the RMS value. Since we aimed to find the parameters that fit most with a training dataset, The average values of the 18 sets of parameters are as follows: $m = 0.1473 \pm 0.0203$, $g = 0.325 \pm 0.0917$, $h_1 = 0.0278 \pm 0.0249$, $h_2 = 0.1667 \pm 0.0236$, $\alpha = 0.0014 \pm 0.0058$, and $\beta = 0.1889 \pm 0.0356$. The box plots in Figure 3 demonstrate the range and interquartile range of parameters, $m$, $g$, $h_1$, $h_2$, $\alpha$, and $\beta$. 

![Variability of parameters among 18 sets of parameters that fit best with a training dataset based on RMS values. $m$ represents the probability of further metastasis along the lymphatic pathway, $g$ represents the probability of primary tumor growth to a higher T stage, $h_1$ and $h_2$ represents the probability of initial metastasis to the two potential first echelon nodes, and $\alpha$ and $\beta$ are coefficients in a linear function that represents the probability of diagnosing a tumor in that specific state.](image-url)
Table 1: Comparison of our model’s predictions with patient data of the Lateral Tongue Tract when $m = 0.15$, $g = 0.4$, $h_1 = 0.05$, $h_2 = 0.2$, $\alpha = 0.0$, and $\beta = 0.175$. The numbers in the table demonstrate the number of patients for each site, one from our model and the other from the patient dataset.

<table>
<thead>
<tr>
<th></th>
<th>Model Data</th>
<th>Model Data</th>
<th>Model Data</th>
<th>Model Data</th>
<th>Model Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T4</td>
<td></td>
</tr>
<tr>
<td>No Metastases</td>
<td>12.6</td>
<td>14</td>
<td>10.5</td>
<td>12</td>
<td>3.8</td>
</tr>
<tr>
<td>Submandibular LC (I)</td>
<td>2</td>
<td>1</td>
<td>3.4</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>Superior cervical LC (II)</td>
<td>4.6</td>
<td>3</td>
<td>8.1</td>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>Middle cervical LC (III)</td>
<td>0.9</td>
<td>2</td>
<td>2.1</td>
<td>5</td>
<td>1.4</td>
</tr>
<tr>
<td>Inferior cervical LC (IV)</td>
<td>0.2</td>
<td>1</td>
<td>0.4</td>
<td>0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 2: Comparison of our model’s predictions with five external datasets, one patient dataset and four literature datasets of the Lateral Tongue Tract when $m = 0.15$, $g = 0.4$, $h_1 = 0.05$, $h_2 = 0.2$, $\alpha = 0.0$, and $\beta = 0.175$. The numbers in the table represent cumulative values across all T stage per nodal echelons.

<table>
<thead>
<tr>
<th></th>
<th>Model Patient</th>
<th>Model Lim*</th>
<th>Model Dias*</th>
<th>Model Sparano*</th>
<th>Model Dogan*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model Data</td>
<td>Model Data</td>
<td>Model Data</td>
<td>Model Data</td>
<td>Model Data</td>
</tr>
<tr>
<td>No Metastases</td>
<td>27.8</td>
<td>28</td>
<td>32.9</td>
<td>15</td>
<td>129.4</td>
</tr>
<tr>
<td>Submandibular LC (I)</td>
<td>7.9</td>
<td>5</td>
<td>7.7</td>
<td>8</td>
<td>33.5</td>
</tr>
<tr>
<td>Superior cervical LC (II)</td>
<td>19</td>
<td>18</td>
<td>17.9</td>
<td>7</td>
<td>78.7</td>
</tr>
<tr>
<td>Middle cervical LC (III)</td>
<td>5.1</td>
<td>12</td>
<td>4.3</td>
<td>6</td>
<td>19.5</td>
</tr>
<tr>
<td>Inferior cervical LC (IV)</td>
<td>1.1</td>
<td>6</td>
<td>0.8</td>
<td>0</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Model validation

The model was validated using data from four previously published original articles. Similar to the internally collected data used to train the model, datasets from published article only contained patients data for their initial presentation and excluded recurrent cancer. External datasets report information on $T$ stage of the patient’s the as well as the anatomical nodal echelon of metastasis. Unlike internally collected data, the previously published data is sparse and incomplete therefore staging and nodal echelon data is reported in aggregate. Thus, the model is tested against the cumulative values across all T stage per nodal echelons and the cumulative values across nodal echelons per T stage. Of note, of the five external datasets, two come from studies that where patients underwent a supra-omohyoid neck dissection and level IV and V information is not available. Given the clinical rationale for a supra-omohyoid neck dissection (extremely rare metastasis to level IV or V) it is reasonable to assume a zero values for all level IV and V metastasis probabilities.

In order to test our model prediction, we produced an output after running our model using the following set of parameters that fit best with a training dataset: $(m, g, h_1, h_2, \alpha, \beta) = (0.15, 0.4, 0.05, 0.2, 0.0, 0.175)$. We obtained the number of patients for each site and compared it with one from the patient datasets as shown in Table 1. For comparing our model predictions with other external datasets, we calculated cosine similarity to measure the similarity between our model output and each external dataset using the Equation 3. The cosine similarity between our model output and the training dataset is closer when the cosine similarity approaches to 1. Given those two vectors, $A$ from our model output and $B$ from a training dataset, the cosine similarity is defined as follows:

$\cos(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$ (3)

Table 2 shows cumulative values across all T stage per nodal echelons from our model output and each external dataset. With the cumulative values, we created two vectors, $A$ and $B$, each of which consists of cumulative values across all T stage per nodal echelons from our model output and external datasets, respectively. We then calculated the cosine similarity between our model output and each external dataset. Table 3 shows the cosine similarity values between our model prediction and each external dataset.

Discussion

We have proposed a novel Markov chain model for prediction of HNSCC tumor dissemination through the lymphatics of the head and neck. In establishing parameters, we focused primarily on a subset of HNSCC, the case in which the primary tumor arises in the lateral tongue region. As such, we were able to generate predictive probabilities using our
Table 3: Cosine similarity of our model’s predictions with one patient and four literature data of the Lateral Tongue Tract when $m = 0.15, g = 0.4, h_1 = 0.05, h_2 = 0.2, \alpha = 0.0, \beta = 0.175$. The values of cosine similarity are given.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cosine Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient data</td>
<td>0.969</td>
</tr>
<tr>
<td>Lim$^4$</td>
<td>0.949</td>
</tr>
<tr>
<td>Dias$^p$</td>
<td>0.967</td>
</tr>
<tr>
<td>Sparano$^4$</td>
<td>0.975</td>
</tr>
<tr>
<td>Dogan$^5$</td>
<td>0.984</td>
</tr>
</tbody>
</table>

model and validate them against known clinical data. Now that the foundations of our model have been established and tested, we can proceed in future work to expand the Markov chain model to include further subsets of HNSCC that arise from different primary sites in the upper aerodigestive tract. There are many such different permutations for metastasis along the lymphatics which we hope to ultimately be able to incorporate into an expanded version of our model. Additionally, we aim to refine our Markov chain model to account for factors such as potential contralateral tumor dissemination. Such factors could be addressed by adding new dimensions onto our existing array. Ultimately, once our Markov chain model is fully expanded, we hope to use it to accurately and precisely predict the extent of micro-metastases dissemination in all cases and subsets of HNSCC, thereby creating a powerful new tool that clinicians can use to provide necessary care to patients while minimizing the deleterious effects of over-irradiation.

**Limitations**

The primary site was chosen due to the relatively large number of patients in our dataset with anterior tongue primary tumors (not including the base of tongue.) Our model only considers ipsilateral metastasis. Contralateral metastasis can occur and we are working to incorporate bilateral metastases into our model. Our model does not include the possibility of skip metastasis. While skip metastases are relatively infrequent, we are investigating ways of incorporating the possibility of skip metastases into future models.

In regard to test and training sets, the literature datasets we used as test sets contain aggregated probabilities for both T stage and lymph node level. We compared the involvement of each nodal group in the test set to what was predicted by the model for the same T stage distribution of that test set. This is a limitation of the existing literature. The patient dataset we used does contain discrete (non aggregated) probabilities that we calculated for each T stage and lymph node level. However, the training set is relatively limited in size, which results in small numbers of observations particularly for more advanced tumors (since most tumors are treated before they reach an advanced stage.)

We assume that $m$ is the same but we could add other parameters that release this assumption. We realize that this is likely an oversimplification and the model may be able to achieve better performance with more variables. Increasing the number of variables with the current amount of training data runs the risk of overfitting. A reasonable assumption is that the probability of tumor metastasizing between different nodal regions may have somewhat different values. However, the probability of tumor metastasizing from one specific nodal region to another specific nodal region is likely independent of the primary site (since all of these tumors are of the same histologic type.) Our future work is to build additional models for different primary sites that share all or part of the same lymphatic drainage pathways. This will allow us to use additional data to fit specific parameters that represent the probability of metastasis between specific lymphatic chains.

We also assume that the probability of growth is a constant parameter. As above, this is likely an oversimplification. However, the TNM tumor stage has been designed to divide each primary tumor site into four “clinically relevant” divisions, which span different primary sites. We therefore felt that this was a reasonable simplification as increasing the number of growth parameters carries the same risk of overfitting described above.

**Acknowledgements**

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R21-LM012075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The first author (Jung) acknowledges support from the Korean Government Scholarship and Mogam Science Scholarship.
References


1839
A Novel Schema to Enhance Data Quality of Patient Safety Event Reports

Hong Kang, PhD, Yang Gong, MD, PhD
School of Biomedical Informatics, the University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

The most important knowledge in the field of patient safety is regarding the prevention and reduction of patient safety events (PSEs). It is believed that PSE reporting systems could be a good resource to share and to learn from previous cases. However, the success of such systems in healthcare is yet to be seen. One reason is that the qualities of most PSE reports are unsatisfactory due to the lack of knowledge output from reporting systems which makes reporters report halfheartedly. In this study, we designed a PSE similarity searching model based on semantic similarity measures, and proposed a novel schema of PSE reporting system which can effectively learn from previous experiences and timely inform the subsequent actions. This system will not only help promote the report qualities but also serve as a knowledge base and education tool to guide healthcare providers in terms of preventing the recurrence of PSEs.

Introduction

Retrospective analysis of health data holds promise to expedite scientific discovery in medicine and constitutes a significant part of clinical research. Informatics technology has helped improve the efficiency by replacing paper-based systems in many healthcare organizations and garnering rich health data. For example, dozens of patient safety event (PSE) reporting systems have been established to enable safety specialists to analyze events, identify underlying factors, and generate actionable knowledge to mitigate risks, such as PSRS (Patient Safety Reporting System) and AHRQ Common Formats (Common Definitions and Reporting Formats). These systems are initiatives to improve patient safety because data supports further learning and actionable knowledge. However, the success of such systems in healthcare is yet to be seen because the data quality could not reach the satisfying level. Therefore, maintaining the quality of healthcare data is widely acknowledged as problematic but critical to effective healthcare.

A PSE is any process, act of omission, or commission that resulted in hazardous health care conditions and/or unintended harm to the patient. Improving the data quality for PSE reporting systems is a challenging task since many factors are involved: inadequate management structures for ensuring complete, timely and accurate reporting of data; inadequate rules, training, and procedural guidelines for those involved in data collection; fragmentation and inconsistencies among the services associated with data collection; and the requirement for new management methods which utilize accurate and relevant data to support the managed care environment. Sometimes reporting systems cannot guarantee qualities because many reporters complete the task halfheartedly. For example, in the AHRQ Common Formats reports, we found more than 80% reporters tend to choose ambiguous options such as “Unknown” which are meaningless for learning rather than explicit “Yes” or “No” options. These systems will become redundant databases as more and more low-quality reports emerge, from which researchers cannot learn any experience from previous cases. Also, there is a lack of connections among different PSE resources due to the diverse report formats. Users have to spend plenty of time manually investigating helpful information from previous cases. Therefore, facing the gap that end-users cannot receive any perceived benefit during and after the reporting, there is an urgent need of a new generation PSE reporting system which can continually enrich the PSE knowledge base by annotating new reports and solutions according to their features, and provide all necessary information (e.g., solutions and prevention options) to the reporters’ cases by measuring the similarities of PSE reports. The quality of the reports will be improved consequently since reporters can receive “benefits” for their reporting rather than passively completing the tasks.

The primary challenge of the new reporting system is measuring the similarity of PSEs, specifically, how to calculate the similarity score between two PSE reports. Different fields provide diverse definitions for the term “similarity”, but there is still no an explicit definition in PSE domain. Based on the theory of cognitive science, such a system handles tasks in the case-based reasoning (CBR) cycle including retrieve, reuse, revise, and retain. Specifically, we firstly measure the similarity of the current problem to previous ones stored in a sort of database with their known solutions, then retrieve one or more similar cases and attempt to reuse the solution of one of the retrieved cases. The solution is assessed and revised before being proposed. At last, the problem description and its
solution is retained as a new case learned by the whole system to solve new problems. Therefore, the measurement should be based on an assumption that similar PSEs have similar solutions.

AHRQ Common Formats\textsuperscript{6} and International Classification of Patient Safety (released by WHO)\textsuperscript{12} defined incident types and event categories for PSE, which are widely accepted and commonly used in patient safety community. Recently, researchers are trying to develop new sets of categories in order to better serve the reporting. Nevertheless, the data formats of individual case were all in the form of ontology (a hierarchical data structure to manage terms and their relationships). Therefore, the comparison between two PSEs could be technically processed through semantic similarity measure as a function that, given two ontology terms or two sets of terms annotating two entities, returns a numerical value reflecting the closeness in meaning between the two\textsuperscript{13}. The semantic similarity algorithms have been generally applied in many fields, such as bioinformatics\textsuperscript{13-15}, geoinformatics\textsuperscript{16}, linguistic\textsuperscript{17} and natural language processing (NLP)\textsuperscript{18, 19}, etc. The Gene Ontology (GO)\textsuperscript{20} is the main focus of investigation of semantic similarity in molecular biology, because not only it is the ontology most commonly adopted by the life sciences community, but also comparing gene products at the functional level is crucial for a variety of applications. Numerous researches have demonstrated that the functional relatedness between genes products with GO annotations can be well measured by semantic similarity algorithms\textsuperscript{13, 21-25}, which provide major significance for gene function studies. Similarly, the patient safety community also need an approach to compare PSEs, then offer the users potential hints of solutions to current cases. Intuitively, the form of PSE seems similar to that of GO, since a number of taxonomies have been designed for labeling case through ontology annotations. Accordingly, the methods that work effectively to compare GO products might be also feasible when identifying similarities in PSEs. However, to our best knowledge, the semantic similarity algorithms have never been adopted and assessed in patient safety area.

In this study, we hypothesize that 1) the similarity between PSEs can be measured if they are annotated by the same ontology/taxonomy in patient safety domain, and 2) similar PSEs have similar solutions. We utilized and assessed the semantic similarity measures on the PSE datasets of AHRQ WebM&M (Morbidity and Mortality Rounds on the Web)\textsuperscript{26} and AHRQ Common Formats\textsuperscript{6}. Based on this model, we proposed a novel schema which can process the comparison tasks for PSEs and provide the reporters pertinent suggestions about solutions and prevention options for their cases. The schema will help develop a new generation of PSE reporting system which could improve the data quality of PSE reports by arousing the enthusiasm and motivation of reporters, and hold promise in preventing the recurrence and serious consequences of PSEs.

Method

Datasets of PSE Reports

AHRQ WebM&M: As a public assessable resource, AHRQ WebM&M Web site\textsuperscript{26} represents illustrative cases of confidentially-reported PSEs on the internet, accompanied by straightforward evidence-based expert commentaries. It also provides a taxonomy with 219 terms to describe the features of PSE from six perspectives: safety target, error type, approach to improving safety, clinical area, target audience and setting of care. Each case in the database has an individual set of annotations on this taxonomy labeled by experts. As of January 7, 2016, 366 cases have been posted on the site. The data size cannot hinder WebM&M from becoming an adequate resource for researching purpose because of the high-quality contents and the diversity of event profiles.

AHRQ Common Formats: AHRQ created the Common Formats to help providers uniformly report PSEs and to improve health care providers’ efforts to eliminate harm\textsuperscript{6}. The Common Formats is a group of standardized questionnaire-based forms with nine subtypes defined by PSO (Patient Safety Organization), including blood or blood product, device or medical/surgical supply including health information technology, fall, healthcare-associated infection, medication or other substance, perinatal, pressure ulcer, surgery or anesthesia, and venous thromboembolism. There is an individual questionnaire for each subtype which contains 10 to 30 single or multiple choice questions, or subjective questions. Missouri Center for Patient Safety (MCPS), a PSO institute and the collaborator of this study, started to collect real PSE reports from several hospitals in Missouri since 2008. So far the database has included more than 41,000 PSE reports. As part of the national level database, the data provided by MCPS can represent all PSE data of PSO program.

Semantic Similarity Measures

Semantic similarity assesses the degree of relatedness between two entities by the similarity in meaning of their annotations. There are two types of semantic similarity approaches when comparing terms, edge-based and node-based (Table 1). Edge-based approaches are based on counting the number of edges in the graph path between two
terms\textsuperscript{27}, for instance, the shortest path or the average of all paths. Correspondingly, node-based approaches focus on comparing the properties of the terms themselves, their ancestors or descendants. Information Content (IC), the most commonly used approach in GO studies, belongs to this category. IC gives a measure of information for every term and makes them comparable. Edge-based and node-based approaches are intended to score the similarity between two terms, and must be extended to compare sets of terms such as gene products and incident cases. Pairwise and groupwise approaches are the two types of strategies for this issue. Every term in the direct annotation set A is compared against every term in the direct annotation set B in pairwise approaches, then the semantic similarity is considered by every pairwise combination of terms from the two sets (average, the maximum, or sum) or only the best-matching pair for each term. Groupwise approaches calculate the similarity directly by set, graph, or vector. Set approaches are not widely used since they only consider the direct annotations that would lose a lot of information; based on set similarity techniques, graph approaches represent entities as the subgraphs of the whole annotations and calculate the similarity using graph matching techniques; vector approaches compact the information in vector space as binary fingerprints which are more convenient for comparison.

Table 1. Pros and Cons of semantic similarity measures

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measures for terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge-based</td>
<td>Intuitive, easy to perform.</td>
<td>Assume all the nodes and edges are uniformly distributed and treat them who are in the same depth equally, which is not applicable for real data.</td>
</tr>
<tr>
<td>Node-based</td>
<td>Measure the terms independent of their depth in the ontology.</td>
<td>The common used term would make more contribution when calculating the similarity.</td>
</tr>
<tr>
<td><strong>Measures for sets of terms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>The contributions from every pair of terms are concerned.</td>
<td>Over-reliance on the quality of data; time-consuming.</td>
</tr>
<tr>
<td>Groupwise</td>
<td>Compare term combinations from a macro view instead of relying on integrating similarity between individual terms; time-saving.</td>
<td>Excessive choices could be a trouble.</td>
</tr>
</tbody>
</table>

We applied and compared three typical semantic similarity measures in this study:

**Information Content (IC).** As a classic node-based approach, IC gives a measure on how specific and informative a term is. Towards PSE reports, it assumes that a term with higher probability of occurrence may contribute less when measuring the similarity. In this study, the pairwise strategy which calculated the similarity for all pairs of terms and assessed them with average score was adopted. And we used Lin’s measure of similarity\textsuperscript{28} which accounts the IC values for each of term \( t_1 \) and \( t_2 \) in addition to the lowest ancestor shared between the two terms.

**Normalized Term Overlap (NTO).** NTO\textsuperscript{21} considers the set of all direct annotations and all of their associated parent terms. Theoretically, NTO might be applicable if the ontology of WebM&M data has been well defined and annotated. The only concern is that the depth of its tree structure is not deep enough which may decrease the comparison resolution. However, in order to further study the applicability of this typical graph-based groupwise approach, we also enrolled it in our assessment.

**Vector Space (VS).** VS compacts the annotations of a set of terms into a binary vector which is more comparable because the model is based on linear algebra with lots of mature algorithms which can measure similarity, such as cosine measure\textsuperscript{29}. Similar to IC, a variation of VS approach has been used in ontology-based similarity. The approach generates a weight for each term based on the frequency of its occurrence in the corpus, and then replaces the non-zero values in the binary vector with these weights. As the WebM&M cases are well annotated in an ontological structure, VS measure, a vector based groupwise approach, may be potentially applicable to measure the similarity between PSEs.

**Expert Review Based on 4-Point Likert Scale Measure**
To each query case, the semantic similarity algorithms ranked all the other cases in the database according to the similarity scores. A group of cases with the same interval were chosen from the rank for expert review. To assess the result given by the algorithms, we adopted a 4-point Likert scale measure which contains 1-irrelevant, 2-somewhat irrelevant, 3-relevant, and 4-highly relevant. Three experts were invited to label every case with one of the four scales according to the degree of similarity between this case and the given query case. After the experts completed the review, two rounds of discussion were opened for them to provide a final review result. If they cannot reach an agreement to certain case, the case would be labeled by a majority. The final expert result was treated as a golden standard. Any case that was labeled by either 1 or 2 in both expert’s result and algorithm’s result was regarded as “agreement” and judged as being irrelevant to the query case; conversely, the one that was labeled either 3 or 4 in both of the two results was also regarded as “agreement” but classified as being relevant to the query. The agreement ratio between final expert review and algorithm (sample agreement ratio) was calculated by dividing the numbers of agreement cases by the number of total cases. Then we randomly labeled the same group of cases for 10,000 times and calculated the agreement ratios respectively (random agreement ratios). To evaluate the performance of certain algorithm of the three, one sample t-test was adopted to examine the mean difference between the sample agreement ratio and the random agreement ratios mean (power analysis). The main steps of the evaluation are shown in Figure 1.

Figure 1. Performance evaluation for the similarity searching model. 1) A 4-point Likert Scale measure which contains 1-irrelevant, 2-somewhat irrelevant, 3-relevant, and 4-highly relevant was adopted; 2) Three experts labeled every case with one of the four scales according to the degree of similarity between this case and the given query case, meanwhile the model ranked the sample cases according to the similarity scores and labeled them with the four scales; 3) The sample agreement ratio was calculated by dividing the numbers of agreement cases by the number of total cases. 4) The performance of the model was evaluated by comparing the sample agreement ratio to the random agreement ratios.

Results

A Similarity Searching Model for PSE Reporting System

To provide solutions for a given PSE report, it should be figured out that which previous cases are similar to the given case. The similarity between PSEs can be measured by comparing their annotations on the same taxonomy. Similar annotations would obtain a high similarity score. AHRQ WebM&M is an appropriate dataset to implement and evaluate the similarity searching model since it provides a taxonomy with 219 terms to describe the features of PSE from six perspectives. In this study, to each pair of the 366 PSE reports, similarity scores were calculated by the three semantic similarity measures (VS, NTO, and IC) separately. As a result, three individual similarity matrices were generated to store all similarity scores.
We developed a local web server to present and evaluate the similarity searching model. Users can either report a new PSE or choose an existing case in the datasets of WebM&M or AHRQ Common Formats as a query case to initiate a similarity searching session. The server will calculate the similarity scores between the query case and every other case, and rank all these cases according to the scores (high to low). Figure 2 shows a screenshot of the similarity searching result when using one of the case in WebM&M named: *Ebola: Are We Ready* as a query. The similar cases calculated by VS method were presented on the right side with case topics, dates, and authors. When any case is selected by the user, its annotation will be presented on the left side. Any click on these terms will trigger a new searching session using the clicked term as a new keyword. The system also provides three action buttons on the right side of each similar case including: 1) *ask for details* (the paper icon), to show the details of this case, as well as the commentaries; 2) *choose as a query* (the magnifier icon), to launch a new similarity searching session using this case as a query; 3) *I agree* (the thumb up icon), to receive user feedback and improve the similarity matrices.

![Figure 2](image.png)

**Figure 2.** Similar cases of the query case “Ebola: Are We Ready?” calculated by the PSE similarity searching model (Vector Space method)

**Model Assessment via Expert Review**

We conducted a detailed study of the three semantic similarity algorithms based on AHRQ WebM&M taxonomy. Case 241, a typical event of nosocomial infections, was chosen as the test query. All the other cases were ranked by our similarity searching model according to the similarity scores, and there were 49 of overall 365 cases carrying nonzero similarity scores in the result list. We randomly extracted 15 cases from the ones with nonzero scores, and 5 cases from the rest, since we supposed the irrelevant cases may occupy 25% of the list according to the natural proportion of the 4-point Likert scale. Three experts with clinical experience reviewed and rated the 20 cases without implication. The result showed that the agreement of the three experts was 90% before the group discussion, and 100% after the first round of discussion. It was even more encouraging that the only two cases judged as “relevant” to the query by the experts have the highest similarity scores calculated by our model.

Further analysis was then processed to figure out the agreement between algorithms and experts. By comparing with random model, the VS model and NTO model reflect a significantly higher consistency with the experts’ review (Table 2). It is feasible and practical to apply semantic similarity model to measure the similarity of PSEs.

<table>
<thead>
<tr>
<th>Table 2. The agreements between algorithms and experts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
</tr>
<tr>
<td>VS model</td>
</tr>
<tr>
<td>NTO model</td>
</tr>
<tr>
<td>Random model</td>
</tr>
</tbody>
</table>

1844
Recommending Solutions to Query Reports

We collected solutions for the most common PSE subtype, patient falls, which occurred about 700,000 to 1,000,000 times each year in U.S. hospitals\(^3\). The solutions were collected and curated from three resources: 1) Pennsylvania Patient Safety Authority Site\(^3\); 2) AHRQ WebM&M commentaries; 3) Staff of a PSO institute, Missouri Center for Patient Safety (via interviews). According to the 13 questions in the fall form of AHRQ Common Formats, the solutions of the fall events were divided into two categories: general solutions (applicable to any fall event) and specific solutions (applicable when certain answer options are chosen). Totally 15 general solution entries and 40 specific solution entries were refined as the solution dataset for fall events. In order to provide a succinct solution recommendation to the query case, we further grouped the general solutions into six subtypes: event reporting, education, patient monitoring, risk assessment, communication, and assistive devices.

---

**Figure 3.** Solution recommendation for a fall event reported in AHRQ Common Formats. The specific solutions are recommended dynamically according to the report options. E.g., the solution entry “Re-evaluate types of assistive devices used by the facility to prevent falls” was presented because the reporter chose “b. Ambulating with assistance and/or with an assistive device or medical equipment” to answer the Question 6.
As shown in Figure 3, when a user completed a fall report in AHRQ Common Formats (case ID 399), the system provided both similar cases and potential solutions for the user to switch at will. All the specific solution entries have been directly linked to the specific answer option(s), which means an entry will be presented only if the associated option(s) is(are) chosen during the reporting. For example, the solution entry “Re-evaluate types of assistive devices used by the facility to prevent falls” was presented because the reporter chose “b. Ambulating with assistance and/or with an assistive device or medical equipment” to answer the question what was the patient doing or trying to do prior to the fall (Question 6). The general solutions were presented for all fall events. In this way, reporters will spontaneously improve their reports since the quality of the reports relates directly to the quality of the specific solutions they received. It will be a win-win situation that reporters really learn something from the system, and the report quality gets improved.

Discussion

In this study, we proposed a novel workflow to improve the data quality of PSE reporting system, which includes seven key modules: data collection and management, algorithm implementation based on semantic similarity measures, expert review, agreement analysis, statistical test, user interface, and user feedback mechanism (as shown in Figure 4). Using fall events as an example, the proposed schema for PSE reporting system can process the comparison tasks and provide the reporters pertinent suggestions about solutions and prevention options for their reporting cases. The schema will be applied in all PSE subtypes in our future work. It will mend the gap that most current PSE reports are in low quality because the end users cannot receive any perceived benefit during and after the reporting. The task of learning and analyzing from previous experiences will be completed by the intelligent system instead of solely relying on healthcare providers. The more detailed users report, the more accurate suggestions they will receive. Such benefits would greatly enhance the quality of PSE reports. Moreover, the semantic similarity module was embedded into the reporting system which allows users to provide feedback such as whether they agree or disagree with certain similar case or solution by clicking a user feedback button. All the feedback will be returned to the algorithm implementation step and help update the weights of related similarity matrices. In this way, the performance of the similarity searching model will be improved dynamically. The model will be progressively stable and enhanced as the involvement through the users’ feedback.

Figure 4. The workflow to improve the data quality of PSE reporting system. Data collection and management, algorithm implementation based on semantic similarity measures, expert review, agreement analysis, statistical test, user interface, and user feedback mechanism are the key modules to improve the data quality of the proposed PSE reporting system.

A challenge of the future work is how to standardize PSE storage and management in order to overcome the lack of interoperability and communication. In this study, we found that PSEs can be compared to each other if they are managed by a hierarchical data structure (e.g., taxonomy) according to their characteristics. Therefore, our future plan is developing a knowledge base in patient safety domain to standardize PSE reports and solutions at the data
level. This standardization will make connections not only among diverse reporting formats but also between solutions and PSE reports. The solutions will be recommended based on the similar cases instead of answer options, which will make the recommendations more targeted. This plan is feasible and low-risk since we are not going to develop a brand-new knowledge base, instead, we will extend part of the terms and relationships defined by International Classification for Patient Safety (ICPS)\textsuperscript{35} to a knowledge base in patient safety domain. The ICPS is carried out under the World Alliance for Patient Safety of the World Health Organization (WHO). ICPS represents the knowledge related to PSE that occur to patients during hospital stays. For the time being, the representation has focused on falls and pressure ulcer incidents. To construct this knowledge base, we plan to extract the sub-classes including terms and relationships about PSE from ICPS. Then we will use Competency Questions (CQ), a classical method for designing and evaluating knowledge bases\textsuperscript{36}, which are interview questions that require candidates to provide real-life examples as the basis of their answers. Two reviewers with background knowledge in patient safety will participate in the CQ development. To each subtype of PSE, they will review a group of cases and generate the CQs. Each CQ must not be duplicated in its meaning with any of the others to the best of reviewers’ judgment. Then we will merge the CQs with the same meanings and break down the subdivisible ones. Thereafter, corresponding classes will be selected and combined with those extracted from ICPS, and a knowledge base describing each PSE subtype will be generated. The developing schema holds promise in improving patient safety and potentially triggering a revolution for data management and analysis in health care industry.

Clinical Decision Support Systems (CDSS) link clinical observations with health knowledge to assist clinical decisions\textsuperscript{37}. The systems influence clinician’s decisions and consequently enhance health-care quality. However, most CDSS suffer from common problems in usability, which have received significant attention in the patient safety community\textsuperscript{38-41}. Several subtypes of PSE are associated with clinical decisions, for example, a medication event may lead a drug-drug interaction. Therefore, in the future work, we will design several specific interfaces to CDSS, embed the key function models of our system to some widely used CDSS, and help detect potential risks at the early stage, thus provide better decision support and eventually improve patient safety.

**Limitations**

**Database** The main defect of AHRQ WebM&M dataset is the small sample size which comprises only 366 PSE reports by January 2016. The selection criteria are unclear, however, based on our observation, the cases in WebM&M may have been chosen as the most typical ones in each category. Taking patient fall as an instance, as the most common event type, it only has four records in WebM&M, which does not mean patient fall is infrequent, but rather indicates the other cases might be similar to the four typical samples and thus were not included by the experts. This is probably the main reason that the algorithm Information Content (IC) has the worst performance on this dataset. IC treats the frequency of each term as an important parameter that high frequency means low information content. This rationale does not apply to this experiment because the frequencies in WebM&M dataset cannot represent the real PSE distribution.

The limitation of AHRQ Common Formats is the low report quality. Many hospitals are using the Common Formats to report their events because of the requirement of PSO, which helps PSO collect a large sum of PSE reports. However, the quality of the reports is unsatisfactory since not all the reporters reported their cases spontaneously. We found that some reporters tended to choose the obscure options rather than the explicit ones which may cost more time to figure out the details. For example, reporters are more likely to choose the option “Unknown” if provided. Unfortunately, most “Unknown” choices are least meaningful for further analysis, and called missing values. This is consistent with our previous finding\textsuperscript{42}.

However, according to the status quo, these two report formats seem to be the best choices for this study since they are by far the only accessible PSE databases with hierarchical feature labels. We believe the quality of reports will be improved if reporters can really learn something from the system during and after their reporting. That is one of the reasons we propose this study.

**Assessment Strategy** All the assessments in this study were processed with the help of health care experts since there is no golden standard for PSE similarity. Each expert likely brings various biases on his/her different perspectives. For example, a clinician may judge the similarity between PSEs by measuring their severity, while a nurse may judge the similarity based on solution finding. The biases cannot be fully avoided. However, we provided more targeted introductions before every round of expert review, and adopted interviews rather than questionnaires in order to help the experts better understand our purpose.

**Conclusion**
Voluntary PSE reporting systems have a great potential for improving patient safety through wide adoption and effective use in healthcare. The similarity analysis of PSE is a key to the success of such systems. Focusing on the quality issue of current reporting systems, this study proposed a novel schema to improve the data quality by gathering information from previous experiences and informing subsequent action in a timely manner. Healthcare providers can learn how to avoid hazardous consequences and prevent their recurrence after they report their PSEs to our system. These benefits encourage the reporters to spontaneously provide reports with higher quality, since the more detailed they report, the more accurate suggestions they will receive. Furthermore, the schema holds promise in developing a new generation of self-learning PSE reporting system based on a standard knowledge base in patient safety domain. As the system being used by more organizations for a longer period of time, the internal knowledge base will become more intelligent, eventually provide better services.

Acknowledgement

We thank Drs. Jing Wang and Nnaemeka Okafor for their expertise and participation in expert review. This project is supported by UTHealth Innovation for Cancer Prevention Research Training Program Post-Doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #RP160015), Agency for Healthcare Research & Quality (1R01HS022895), and University of Texas System Grants Program (#156374).

References

Public Health Data for Individual Patient Care: Mapping Poison Control Center Data to the C-CDA Consultation Note

Aly Khalifa, Guilherme Del Fiol, MD, PhD, Mollie R. Cummins, PhD, RN, FAAN
University of Utah, Salt Lake City, UT, USA

Abstract

We are developing a new process of health information exchange supported collaboration that leverages the HL7 consolidated CDA standard through four document types (consultation note, history and physical, progress note and discharge summary). The focus of the present study is the C-CDA consultation note template that will be submitted from poison control centers (PCCs) to emergency departments (EDs) with treatment recommendations. Specifically, we aimed to (i) create computable mappings between a poison control center database and the C-CDA consultation note template; and (ii) assess the extent and nature of information types that successfully map to discrete data elements in a poison control center database. The resulting template and mappings can be used to implement standards-based health information exchange between PCCs and EDs in the U.S. This is a part of the first formal effort to leverage health information exchange standards to support PCC-ED collaboration.

Introduction

Poison Control Centers (PCCs) are responsible for handling poisoning cases all over the U.S.1. PCCs receive telephone calls from both the general public and healthcare facilities2, especially emergency departments (EDs). EDs ask for the consultations from PCCs via an iterative telephone call process, that carries a lot of inefficiencies and vulnerabilities that may greatly affect patient safety and the provided healthcare quality3. Therefore, electronic health information exchange (HIE) between EDs and PCCs is a promising and complementary alternative to the telephone-based approach4. The overall goal of our research is to enable standards-based HIE between PCCs and EDs to improve the efficiency and quality of care for poisoning cases.

Most PCCs use information systems that allow PCC staff to record information about and keep track of active poisoning cases. Although PCC information systems allow the Specialists in Poisoning Information (SPIs) to capture detailed information about poisoning cases, current PCC systems do not support standards-compliant HIE or implement document standards, such as the Health Level Seven (HL7) Consolidated Clinical Document Architecture (C-CDA)5. In addition, existing HIE standards provide no specific guidance to support PCC-ED information exchange. In previous studies, we developed a model for PCC-ED HIE, including: (i) identification and analysis of information types exchanged between PCCs and EDs via telephone calls6; and (ii) mappings between PCC information types and C-CDA documents, sufficient to support communication and information requirements in a poisoning care episode (i.e., referral request, consultation note, progress report, discharge summary)5,6.

The focus of the present study is the C-CDA consultation note that will be transmitted from PCCs to EDs. This document contains essential patient information along with poisoning case information and initial treatment recommendations. Specifically, we aimed to (i) create computable mappings between data elements in the Utah PCC system, an instance of the widely used PCC information system toxiCall67, and the C-CDA consultation note specifications; and (ii) assess the percentage of poisoning communication information types that could be successfully mapped to discrete data elements in toxiCall67. The resulting mappings can be used to guide the implementation of standards-based HIE for PCCs and EDs across the US. This is a part of the first formal effort to leverage HIE standards to support PCC-ED communication process.

Background and Significance

PCCs and their information systems. There are 55 PCCs that serve all 50 U.S. states and territories, 24 hours per day and 7 days per week1. The PCC staff who consult with the public and ED care providers on poison exposures are the SPIs, typically licensed pharmacists or nurses, and medical or clinical toxicologists. The American Association of Poison Control Centers (AAPCC) owns and manages the National Poison Data System (NPDS), to which all U.S. PCCs regularly upload a set of standard data elements describing human poison exposure cases8. These data are used
for epidemiology and surveillance. AAPCC produces an annual report on the aggregate statistics of NPDS data, in addition to providing specialized reports to national agencies\(^1\,^9\). A PCC uses one of four specialized information systems to support the tasks of data entry, retrieval and data transmission to the NPDS\(^10\).

**PCC-ED health information exchange.** Although the management of poisoning cases is an information and communication intensive process that can benefit from HIE, little has been done to enable HIE between PCCs and EDs. Previously, we identified important legal, operational, and clinical considerations in relation to PCC-ED HIE\(^4\). We also analyzed telephone conversations between poisoning specialists and ED staff to (i) identify inefficiencies and vulnerabilities\(^3\); and (ii) to identify and analyze the information requirements for ED-PCC collaboration\(^6\). In a current study funded by the Agency for Healthcare Research and Quality (AHRQ), our goal is “to develop, implement, and evaluate a replicable, scalable infrastructure for HIE supported ED–PCC collaboration”\(^11\). The project started in September 2013 and is expected to complete in September 2018.

**PCC-ED HIE and the C-CDA standard.** C-CDA is an HL7 messaging standard that specifies a common structure and content for electronic clinical documents, according to document type, e.g., history and physical note vs. discharge summary. The ultimate goal of the C-CDA standard is to make a document that is both human and machine readable by combining narrative blocks with coded entries\(^12\). Extensible Markup Language (XML) is used to represent the document content in a structured, computable way. Viewed through a styling sheet (i.e., XSL file), this XML file is formatted for human reading, without visible XML markups\(^13\).

Generally, a C-CDA document is divided into two main parts, the document header and the document body. The C-CDA header contains metadata about the healthcare encounter and information necessary for document housekeeping and management. Specifically, the C-CDA header includes patient demographics, document author(s), document recipient, document type, patient encounter, and document type. The document body represents the clinically relevant information, structured within a set of sections that are specific according to the C-CDA document type.

The C-CDA standard is an implementation guide that includes specifications for 13 C-CDA document types\(^12\). The specification provides a library of templates that can be used to support various HIE use cases. The detailed constraints on document and section levels are provided within the C-CDA Implementation Guide\(^12\). In a previous study, we linked our proposed PCC-ED information exchange events to four C-CDA document types; and mapped poisoning information types identified in previous work to specific C-CDA sections. The four identified C-CDA document types are Consultation, History & Physical, Progress Note and Discharge Summary\(^5\). The present study builds on our previous work by providing physical mappings between a PCC information system (toxiCall\(^6\)) and the C-CDA Consultation Note document type, which is a necessary step towards the interoperability between PCC and ED information systems.

**Methods**

The approach used in this study consisted of three main steps. First, we developed consultation note mockups and sample documents containing the required information content, as determined in previous studies\(^5\,^6\). Second, we developed a computable mapping between the C-CDA schema (conforming to consultation note constrains) and the Utah PCC database, an instance of toxiCall\(^6\). Last, we assessed the percentage and nature of information types that successfully mapped from the PCC database to the C-CDA. An overview of the steps and their sequence is depicted below (Figure 1) where each step is labelled by a letter and these letters are used in the following paragraphs.

**Mockups and sample document.** Mappings between the information types and consultation note sections were previously established (Step - A)\(^5\). We elicited preferences for information content and presentation format from emergency medicine physicians at our collaborating sites. Based on the mappings and physician preferences, we developed two sample mockups using HotGloo prototyping tool (Step - B)\(^14\). The mockups were then validated by SPIs, ED clinicians and project members to ensure that the content of the consultation note meets their needs. We collaborated with these users to make iterative changes to the design, then manually developed a sample consultation note XML file, and validated that file according to the C-CDA R2.1 Implementation Guide and its schema (Step - C)\(^12\). This sample document was valuable as it identified the actual location of data within the document XML structure, especially for the document header part that was not represented within the initial mapping\(^5\).
PCC database and computational mapping. The Utah PCC provided us with access to a copy of their database, an instance of toxiCall®. Filled with a small set of anonymized patient data. The data were anonymized by replacing all patient demographics, SPI related information and identifiers with fictitious information. Therefore, we identified tables that contain the required information for the C-CDA Consultation Note (Step - D); and created mappings between the PCC database and C-CDA Release 2.1 XML schema (Step – E)12, with the permission of Computer Automation Systems, Inc7. We created the mappings using Altova MapForce® version 2016, which allows mappings between different data sources and targets through a graphical user interface15. The mapping process consisted of “dragging and dropping” toxiCall® database fields to the corresponding elements and attributes in the C-CDA Release 2.1 XML schema. We used custom data transformations to map data from multiple source tables and columns to a single element within the XML structure and to represent them in a proper format.

Once the mapping was completed, we used Altova MapForce® to automatically generate consultation notes for the anonymized case records. We validated these instances against the C-CDA schema and the NIST online validation toolkit, (Step - F)16. Also, a human validation was done by the study co-authors and collaborators from the Utah PCC and EDs (Step – G). For human validation, we transformed the documents into HTML using the Extensible Stylesheet Language (XSL) style sheet provided with the C-CDA R2 Implementation Guide12. Once validated, we used Altova MapForce® to automatically generate source code for the mappings in the Java programming language. The Java code was then wrapped with a RESTful Web services layer to be integrated with other systems.

Extent of structural mapping. For comparison with previous efforts, we calculated the percentage of information types that successfully mapped to the C-CDA from the Utah PCC database. For this calculation, the information type was considered to be present when it was available in a discrete (non-narrative) form within the PCC database, and not present otherwise. For data found to be not present, we identified potential alternative sources such as poisoning knowledge bases and additional data entry fields to be collected by SPIs. Details on alternative sources are subject for a future study. The information content to be used in the C-CDA header was not considered in this analysis as the focus was on clinical information requirements represented in the C-CDA body.

Figure 1. Mapping of PCC database fields to CDA R2, in the context of C-CDA consultation note development
Results

Mockups and sample document

Figure 2 shows a sample mockup of a PCC consultation note for a poisoning exposure due to snake bite. This mock-up includes additional HTML formatting within sections, to implement ED user preferences related to information presentation. Figure 3 shows the manually developed XML consultation note, rendered in HTML format. The actual XML file is available at (https://github.com/alykhalifa/PCC-ED-C-CDA-Consultation-Notes). The header section contains metadata such as patient demographics, document author, intended recipient and encounter information (e.g., data and time of the call). The body contains 8 sections: (i) reason for referral; (ii) chief complaint; (iii) history of present illness; (iv) general status; (v) assessment; (vi) plan of care; (vii) past medical history; and (viii) medications. The history of present illness section contains information about the poisoning exposure such as poison information and observed symptoms. The assessment section consists of two main parts. The first part contains patient-specific information, while the second part contains general information about the poisoning exposure itself, such as its toxic dose, common effects and the time to peak concentration. The plan of care section consists of three main parts: (i) recommended diagnostic testing and monitoring; (ii) recommended observation time; and (iii) supportive care. The other sections contained standard information as defined in the C-CDA standard. The sample XML document was successfully validated with no errors using the C-CDA R2.1 schema and the NIST online validation toolkit16.

![Poisoning Consultation Note](image)

---

**Figure 2.** Mockup of a consultation note for a poisoning due to snake bite.

---
Figure 3. Sample C-CDA Consultation Note rendered in HTML format using the XSL style sheet provided with the C-CDA R2.1 standard specification.

**PCC database and computational mappings.**

Mappings between the C-CDA standard and the Utah PCC database, an instance of toxiCall® involved 8 tables and over 250 columns in the PCC database. Out of 38 information types needed to create a C-CDA compliant consultation note that supports poisoning communication, 21 (55%) successfully mapped to discrete fields within the PCC database. Seventeen information types did not successfully map. We defined over 130 constant variables within Altova Mapforce® to support required elements and attributes according to the Consultation Note C-CDA R2.1 constraints.

Table 1 shows the information types that successfully mapped, along with associated data transformations, where “D” means retrieval of code label from the code dictionary table, and “MC” means collection of required information from multiple columns in the PCC database. Table 2 shows the information types that did not successfully map. A snapshot of the auto-generated consultation note is shown on Figure 4. The validation of the generated documents against the C-CDA R2 schema and through the NIST online toolkit showed no errors or warnings, and was validated by domain experts, with only minor, iterative adjustments. The actual mapping file and the generated Java code are available upon request, with the permission of Computer Automation Systems, Inc., as they contain some information proprietary for toxiCall®, while the generated sample consultation notes are available at (https://github.com/alykhalifa/PCC-ED-C-CDA-Consultation-Notes).
One of the challenges faced during the mapping process is that data for a single information type may be present over multiple tables and columns. For example, recommended therapies and interventions are scattered over 68 columns, where each column represents a single therapy/intervention that may be set to certain numeric values if recommended, or left as null otherwise. A user-defined transformation function was developed using Altova MapForce® to identify columns with numeric values, parse their column name (i.e., the treatment name), and instantiate a corresponding treatment instance in the C-CDA R2.1 format.

Discussion

In this study, we mapped a large proportion of the information content necessary to support poisoning communication from a PCC database to the HL7 C-CDA consultation note. The mappings are an important step towards enabling standards-based HIE between a proprietary system (i.e., toxiCall®) used by most PCCs in the US and various EHRs available at emergency departments. These mappings serve two main functions (i) identify existing data sources to automatically populate the consultation note, and (ii) identify informational gaps that need alternative input sources (e.g., data entry fields and/or a poison knowledge base). The Utah PCC information system (like most) was primarily designed to support the public health mission of poison control centers, namely case documentation and collection of NPDS data elements for the purpose of surveillance and epidemiology. PCC information systems are not designed for interoperability or communication with healthcare facilities. Therefore, much of the information needed by the ED clinicians is either orally conveyed through telephone calls and/or documented within a combined narrative field of the PCC database.
Table 1. Information types that were successfully mapped to the Toxicall® database and their data transformations

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Description (Transformations if present*)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exposure information</strong></td>
<td></td>
</tr>
<tr>
<td>Exposure type</td>
<td>General description of the poisoning case, e.g., unintentional food poisoning (D)</td>
</tr>
<tr>
<td>Certainty of formulation</td>
<td>Certainty level of reported poison substance formulation (D)</td>
</tr>
<tr>
<td>Chronicity</td>
<td>Exposure duration and repetition of the poisoning incident (D)</td>
</tr>
<tr>
<td>Establishing background/certainty</td>
<td>More details about the poisoning case such as symptoms, routes of exposure and exposure site (D)</td>
</tr>
<tr>
<td>Substance class</td>
<td>General category of the poisonous substance/product, e.g., sedative (D)</td>
</tr>
<tr>
<td>Substance name (generic)</td>
<td>Generic name of the poisonous substance/product</td>
</tr>
<tr>
<td>Substance amount</td>
<td>Quantity and units for the substance amount and concentration (D, MC)</td>
</tr>
<tr>
<td>Substance name (brand)</td>
<td>Brand name of the poisonous substance/product</td>
</tr>
<tr>
<td>Substance description</td>
<td>More details about the poisonous substance</td>
</tr>
<tr>
<td>Substance form, formulation and type</td>
<td>Physical form of the poisonous substance, e.g., gas and tablet, and more details about pharmaceutical nature, e.g., sustained release (D)</td>
</tr>
<tr>
<td>Substance–non-pharmacological</td>
<td>Names and descriptions for non-pharmacological substances/products</td>
</tr>
<tr>
<td>Time since ingestion</td>
<td>Duration since the time of poison exposure till the documentation/reporting time</td>
</tr>
<tr>
<td><strong>Subjective and objective information</strong></td>
<td></td>
</tr>
<tr>
<td>Chief complaint/reason for visit</td>
<td>Explaining what is the patient suffers from and she/he need to visit an ED (D)</td>
</tr>
<tr>
<td>Caller reported symptoms</td>
<td>Symptoms mentioned by the one who reported for a poisoning incident over a telephone call (MC)</td>
</tr>
<tr>
<td>Unrelated symptoms</td>
<td>Symptoms reported and deemed unrelated to the current poisoning incident</td>
</tr>
<tr>
<td><strong>PCC recommendations and toxicology information</strong></td>
<td>Therapeutic advice on how to handle the poisoning case, in terms of therapies, labs and patient monitoring time (MC)</td>
</tr>
<tr>
<td><strong>ED treatment/management information</strong></td>
<td></td>
</tr>
<tr>
<td>Diagnostic test results</td>
<td>Results of diagnostic test reported to PCC and stored within its database (MC)</td>
</tr>
<tr>
<td>Time laboratory test was performed/drawn</td>
<td>Reported time of conducting lab tests</td>
</tr>
<tr>
<td>Treatment performed (or recommended by the PCC)</td>
<td>Differentiating between treatments that were recommended, performed or both (MC)</td>
</tr>
</tbody>
</table>

*D, retrieve code label from code dictionary table in PCC database; MC, retrieve values from multiple columns in the PCC database
Table 2. Information types that were unsuccessfully mapped to the Toxicall® database.

<table>
<thead>
<tr>
<th>Information Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure information</td>
<td></td>
</tr>
<tr>
<td>Substance information</td>
<td>Pharmacology and pharmacokinetics information such as time to peak concentration and biological pathways</td>
</tr>
<tr>
<td>Substance identification rationale</td>
<td>Description of how the SPI identify the poisoning substance/product</td>
</tr>
<tr>
<td>Patient health history</td>
<td></td>
</tr>
<tr>
<td>Medical history</td>
<td>Patient medical history that may be relevant to the poison exposure and treatment</td>
</tr>
<tr>
<td>Patient medications</td>
<td>Current and past medications of the patient</td>
</tr>
<tr>
<td>Subjective and objective information</td>
<td></td>
</tr>
<tr>
<td>Absence of clinical effects</td>
<td>Absence of some of the expected symptoms for the poisoning incidence</td>
</tr>
<tr>
<td>Mental status</td>
<td>E.g., dizziness, hallucination…</td>
</tr>
<tr>
<td>Physical exam findings</td>
<td>Clinical findings reported to the PCC and stored within its database</td>
</tr>
<tr>
<td>Vital signs</td>
<td>Times and measures of reported vital signs</td>
</tr>
<tr>
<td>PCC recommendations and toxicology information</td>
<td></td>
</tr>
<tr>
<td>Toxic dose</td>
<td>Toxicity threshold for the poisoning substance/product</td>
</tr>
<tr>
<td>Toxicity levels</td>
<td>Toxicity level of the poisoning substance for this specific poisoning incidence</td>
</tr>
<tr>
<td>Clinical effects of substance</td>
<td>Expected clinical symptoms for the poisoning substance/product</td>
</tr>
<tr>
<td>Worst case scenario</td>
<td>The risks and worst clinical effects that may occur in relation to the poisoning incidence</td>
</tr>
<tr>
<td>ED treatment/management information</td>
<td></td>
</tr>
<tr>
<td>Confirmation that treatment was given</td>
<td>Indication if PCC recommendation were applied to the patient at the ED</td>
</tr>
<tr>
<td>Patient discharge medications</td>
<td>Discharge medications prescribed by the ED clinicians</td>
</tr>
<tr>
<td>Patient status</td>
<td>Patient status at the ED for certain date and time</td>
</tr>
<tr>
<td>Plan of care</td>
<td>Plan of care designed by ED clinicians</td>
</tr>
<tr>
<td>Time next laboratory tests will be ordered</td>
<td>Future laboratory tests timing as specified by ED clinicians</td>
</tr>
</tbody>
</table>

This clinically relevant information includes the expected common effects of the poison, time to peak concentration of the poison, more detailed plan of care, history of past illness, and past/current medications taken by the patient. Also, some information currently collected by this PCC lacks adequate detail of structure for the C-CDA. For example, patient age is stored within the PCC database, but the C-CDA requires date of birth. In anticipation of HIE, the Utah PCC has managed to solve these issues via workarounds. For example, date of birth is being collected in a generic PCC database field. However, a more standard approach to collecting information essential to HIE processes, such as date of birth, is preferred. Furthermore, some of the successfully mapped information types (e.g., clinical effects) may need to be presented in a more expressive form in order to be sufficiently meaningful for clinical use. In this case, some potential alternative sources (e.g., additional data entry fields and a poison knowledge base) may be used to provide this intended level of expressiveness, which is the same approach proposed for handling unsuccessfully mapped information types. As a solution, we are currently developing software that enables poison centers to create and send CCDAs using both template narrative data and coded data contained in the Utah PCC database. That software is currently under development and testing.

The PCC database consists almost entirely of NPDS data elements, data elements collected by all U.S. PCCs, and all of the database fields that successfully mapped to the C-CDA represent NPDS data elements. Given the public health oriented nature of PCC data, it is not surprising that many of the information types necessary to support communication
and collaboration in the context of a poison exposure cannot be mapped. However, the capability to partially construct a C-CDA document using the PCC database greatly lowers the manual effort required for data entry on the part of PCC specialists. This would save a lot of the time needed by the SPI to generate the consultation note, and ensure its validity and accuracy, directly supporting rapid communication in these emergency situations and improving usability for the SPIs. Moreover, since the mappings correspond entirely to NPDS data elements, they may potentially be adapted for use by any U.S. poison control center. The Utah PCC database is a local instance of one vendor solution for poison control centers, toxiCall®. U.S. PCCs use one of four different information systems. However, all U.S. poison control center information systems collect NPDS data elements and support the same essential clinical workflow/domain, and so there is some de facto standardization and similarity among different poison control center information systems. The AAPCC, the owner of NPDS, specifies the required data elements and has the authority to add or modify data elements. However, the purpose of NPDS is to support surveillance and epidemiology, and AAPCC hasn’t yet addressed interoperability in support of direct patient care.

This study has three main strengths that support generalizability of the results. First, we involved the intended users of the consultation note during design, development, and validation. Second, we validated the generated consultation notes using both human (i.e., intended users) and machine approaches. Third, the PCC database is an instance of toxiCall®, a widely used system among U.S. PCCs, and all mappings involved NPDS data elements. All U.S. PCC systems are NPDS complaint, so the developed mapping can be implemented or adapted for implementation at most PCCs with minor adjustments, providing an important building block for a nationwide PCC-ED HIE. A limitation of the mapping is the absence of coded entries in the generated consultation note. We conducted the mapping using the CDA level 2 conformance, which does not specify these coded entries, and therefore limits computation for purposes such as clinical decision support and analytics. Nevertheless, the mappings are a critical step towards enabling HIE between PCCs and EDs to support the care of patients exposed to poisoning.

The current telephone-based ED-PCC collaboration process contains safety vulnerabilities and inefficiencies and there is an evident need for better access to PCC consultation information by other healthcare providers, which is not limited to US PCCs. U.S. PCCs are not alone in facing challenges related to interoperability. An identified top barrier for achieving meaningful use stage 3 criteria among primary care practices is suboptimal HIE capabilities. Where a main challenge for realizing effective HIE is the lagging effort in deploying HIE standards. This project created an important building block to achieving HIE for U.S. PCCs by bridging a gap between PCC data and existing HIE standards. Ongoing efforts are now focused on: (i) mapping free-text entries and NPDS codes to standard terminologies (e.g., SNOMED-CT, RxNorm); and (ii) integrating the generated mapping as a module within software that supports the creation, sending, and receipt of HL7 C-CDA documents.

**Conclusion**

As a step toward standards-based HIE between PCCs and EDs, we mapped fields within a PCC database to corresponding fields of the C-CDA consultation note, according to previously defined information requirements. Much of the data was readily available in a discrete form within the PCC database, while other data were available only as part of a combined narrative field, or not at all. We validated the generated notes for their conformance to the C-CDA R2 standard and fulfillment of ED and PCC clinical information needs. The resulting mappings are an important step towards national adoption of HIE between PCCs and EDs. Ongoing efforts include integration of the mappings as a module within software that supports the creation, sending, and receipt of HL7 C-CDA documents.

**Acknowledgements.** This study was funded by the U.S. Agency for Healthcare Research & Quality (4R01HS021472-03). Clinicians from Intermountain Healthcare and the Utah Poison Control Center provided assistance. Thanks to Computer Automation Systems, Inc. for permission to create a mapping to the Utah PCC instance of toxiCall®.
References

Improving risk prediction for depression via Elastic Net regression - Results from Korea National Health Insurance Services Data

Min-hyung Kim, MD¹, Samprit Banerjee, PhD¹, Sang Min Park, MD, PhD, MPH², Jyotishman Pathak, PhD¹
¹Department of Healthcare Policy & Research, Weill Cornell Medical College, New York, NY, USA
²Department of Family Medicine, Seoul National University College of Medicine, Seoul, Korea

Abstract

Depression, despite its high prevalence, remains severely under-diagnosed across the healthcare system. This demands the development of data-driven approaches that can help screen patients who are at a high risk of depression. In this work, we develop depression risk prediction models that incorporate disease co-morbidities using logistic regression with Elastic Net. Using data from the one million twelve-year longitudinal cohort from Korean National Health Insurance Services (KNHIS), our model achieved an Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) of 0.7818, compared to a traditional logistic regression model without co-morbidity analysis (AUC of 0.6992). We also showed co-morbidity adjusted Odds Ratios (ORs), which may be more accurate independent estimate of each predictor variable. In conclusion, inclusion of co-morbidity analysis improved the performance of depression risk prediction models.

Key words: Depression, Risk Prediction Model, Co-morbidity, Korea National Health Insurance Services Longitudinal Cohort Data, Chronic Conditions Data Warehouse (CCW) Condition Algorithms, Logistic Regression, Least Absolute Shrinkage And Selection Operator (LASSO), Elastic Net

Introduction

Depression is a highly prevalent disease with a large societal burden. Major depressive disorder has the one-year prevalence of 6%, and the lifetime prevalence of 17%³, while persistent depressive disorder (dysthymia) has the one-year prevalence of 2%, and the lifetime prevalence of 3%⁴. The estimated societal burden of unipolar depression was 83 billion dollars per year in the US alone in 2007⁵.

However, despite this burden, depression is under-diagnosed at large across the healthcare system in all care settings. A meta-analysis in 2009 concluded that the weighted sensitivity of primary care physicians’ diagnosis on depression was only about half (41.3-59.0%) without the assistance of screening tools⁶. This lead to the under-diagnosis or delayed diagnosis of depression, because many of depressed patients initially present with somatic symptoms to the primary care clinics. In general, 69-73% of depression patients presented to their primary care physicians with somatic symptoms, such as pain, fatigue, and sleep problems⁷.

Data-driven risk prediction models can be beneficial by rapidly classifying high-risk patients who need further evaluation. Risk prediction models can be implemented on Electronic Health Record system (EHRs) in order to provide high-risk alert, as well as clinical decision support. Risk prediction models can also be implemented in health insurance claims data in order to classify high-risk patients, and can be used for accountable care strategy⁸.

Previous work on the prediction modelling of depression include a regression-based depression risk prediction model based on Electronic Health Record data, developed at Stanford University, which reported an area under the receiver operating characteristic (AUROC) of 0.80 for current classification, 0.712 for 6-month prediction, and 0.701 for 12-month prediction⁹. Another work of the depression risk prediction model based on clinical trial data, developed at University of Southern California, reported to have a current classification with an AUROC of 0.81, as well as a sensitivity of 0.65 and a specificity of 0.81 at the institution’s optimized threshold⁹.

However, both these approaches did not explicitly apply co-morbid medical conditions as independent predictors in the depression prediction model. Many medical conditions can affect depression, and depression can also affect certain medical conditions¹⁰. Therefore, application of co-morbidity analysis can improve the performance of the risk prediction models.

Hence, our main hypothesis and the research question to be addressed in this study was whether the co-morbidity analysis can improve the performance of prediction models for depression risk. Our preliminary results indicate that the inclusion of co-morbidity analysis improved the performance of depression risk prediction model.
Study Setting and Data

In this study, co-morbidity analysis and risk prediction modeling was made from one million twelve-year longitudinal data from Korea National Health Insurance Services (KNHIS)\textsuperscript{11}. The sample cohort (N= 1,025,340) was established in 2002 from 2.2% of 46,605,433 individuals from the National Health Information Database, in order to provide public health researchers and policy makers with representative information regarding the utilization of health insurance and health examinations\textsuperscript{12}. The data include demographic profile, health insurance claims data (including in-patient, out-patient, and pharmacy claims), death registry, disability registry, and national health check-up data. With the combination of 18 age groups, 2 genders, and 41 income groups, total 1476 strata were undergone systematic stratified random sampling with proportional allocation\textsuperscript{13} within each stratum, using the individual’s total annual medical expenses as a target variable. During the follow-up years, annual drop-out by death was 0.5% (ranging from 4,929 to 5,229). Each year, a representative sample of newborns (ranging from 7,782 to 9,581), sampled across 82 strata (2 for gender, 41 for parents’ income group), was added to ensure the representativeness of the data.

The diagnosis codes in KNHIS are based on the Korean Classification of Diseases, Sixth Revision (KCD-6), which is compatible with International Classification of Diseases, Tenth Revision (ICD-10). These diagnoses were classified with Chronic Conditions Data Warehouse (CCW) Condition Algorithms (rev. 01/2016) by Centers for Medicare & Medicaid Services (CMS)\textsuperscript{14}. The CCW condition category algorithms are claims-based algorithms to indicate whether treatment for the condition appears to have taken place, which include 27 chronic condition categories and 33 other chronic or potentially disabling conditions categories. Table 1 shows the ICD-10 codes of the depressive disorder, bipolar disorder, schizophrenia in the CMS-CCW algorithm. The ICD-10 codes for depressive disorder, bipolar disorder, schizophrenia were used in the operational definition of the depression case group in this study. The study subjects had two or more encounters with depression diagnosis codes, but less than two encounters with either bipolar or schizophrenia diagnosis codes. The inclusion and exclusion criteria is based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)\textsuperscript{15}.

<table>
<thead>
<tr>
<th>CMS-CCW Conditions</th>
<th>Valid ICD 10 Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressive Disorders</td>
<td>F32.0, F32.1, F32.2, F32.3, F32.4, F32.5, F32.6, F32.9, F33.0, F33.1, F33.2, F33.3, F33.40, F33.41, F33.42, F33.9, F34.1, F34.9</td>
</tr>
</tbody>
</table>

The univariate and bivariate statistics of selected demographic and co-morbidity variables between the depression case group (N=28,256) and complement comparison group (N=1,085,400), based on the operational definition of the case group described in the method section, are shown in Table 2. The case group showed significantly higher percentage of females (68.1%), age (mean 48, standard deviation 19), income decile (mean 5.8, standard deviation 2.5), limb disability (3.3%), neurologic disability (0.9%), visual disability (0.7%), hearing disability (0.6%), but showed significantly lower percentage of social security beneficiaries (1.0%). The case group showed no significant difference in the percentage of residents in Seoul metropolitan area, and cognitive disability. Most of the co-morbidity variables showed statistically significant difference between the case group and comparison group, except cerebral palsy (p = 0.121). The most noticeable difference in the co-morbidity by the ratio of percentage (twelve-year prevalence) was personality disorder (0.9% vs 0.1%), followed by anxiety disorder (31.8% vs 4.9%), dementia & Alzheimer’s disease (2.0% vs 0.4%), and osteoporosis (5.2% vs 1.3%).

Analytic Approach

The operational definition of diagnosis of depression was analyzed in a logistic regression model with socio-economic and co-morbid predictors. Among the available socio-economic variables and co-morbid conditions in KNHIS data, variables for the final logistic regression model was selected with Elastic Net\textsuperscript{16}. The performance of the final logistic regression model with co-morbidity analysis was compared with that of the traditional logistic regression model without co-morbidity analysis.
When the number of predictors is large compared to the sample size, traditional variable selection methodologies may have poor prediction performance for external datasets by overfitting random error or noise, and it has been criticized that the goodness of fit\textsuperscript{17}, significance\textsuperscript{18}, and degrees of freedom\textsuperscript{19} do not reflect the reality. In order to overcome this problem, regularization and shrinkage methods for regression have been developed\textsuperscript{20}. Elastic Net is a regularization method for regression and classification models which compromises the Least Absolute Shrinkage And Selection Operator (LASSO) penalty (L\(_1\)) and the ridge penalty (L\(_2\))\textsuperscript{16}. The LASSO (L\(_1\)) penalty function performs variable selection and dimension reduction by shrinking coefficients, while the ridge (L\(_2\)) penalty function shrinks the coefficients of correlated variables toward their average. The overall Elastic Net is a function of parameters \(\lambda\) and \(\alpha\) (\(0 \leq \alpha \leq 1\)), where \(\lambda\) being a parameter for the level of penalty, while \(\alpha\) being the weight of L\(_1\) penalty and \((1 - \alpha)\) being that of L\(_2\) penalty function. Hence, in this work, we performed variable selection and penalization of collinear predictors by Elastic Net for developing the final logistic regression model.

A robust way to determine the best combination of \(\lambda\) and \(\alpha\) is via a k-fold cross-validation. For the validation of the predictive model, 10\% of the data (\(N = 111,366\)) was set aside as a test data, and 90\% of the data was used as a training data (\(N = 1,002,290\)). We used 10-fold cross-validation on training data, where total observations of the dataset are randomly divided into 10 folds, or partitions. One of the 10 folds is reserved as the internal validation data (\(N = 100,229\)), and the rest of the folds consist the internal training data (\(N = 902,061\)), where statistical models are fitted. After fitting the models, or calculating the coefficients, the models are validated against the reserved fold. This overall process is iterated (repeated) 10 times, so that each fold can be a validation set. This is a preferred method especially when the prediction models need to perform prediction for external datasets, that is, outside of the overall dataset used in the research.

The variables for the traditional logistic regression model without co-morbidity analysis was driven by performing the stepwise backward selection using Akaike’s Information Criterion (AIC)\textsuperscript{21}. The selected variables for the traditional logistic regression model include sex, age, income decile, and disability registration. The variable selection for the final logistic regression model was applied with Elastic Net from the training data, as described above. The selected value of \(\alpha\) was 0.75, and the optimized values of \(\lambda\) was 0.001390648 (\(\log(\lambda)\) -6.577986), although other \(\alpha\) values, including 0.25, 0.5, and 1, did not change the results much. The plot obtained from cross-validation of Elastic Net, showing the change of the Area Under the Curve (AUC) of ROC with different \(\lambda\) (in log scale) for a model assuming an \(\alpha\) of 0.75, is shown in Figure 1. This gives a minimum of 28 variables needed for building an optimized model. Two more variables, acute myocardial infarction and dementia, were added to the final model, because even though those conditions were separated by the CMS-CCW algorithm, the conditions were in spectrum with ischemic heart disease and Alzheimer’s disease, respectively. Therefore, 30 variables were selected for the final logistic regression model. These variables include: sex, age, income decile, acquired hypothyroidism, acute myocardial infarction, Attention Deficit Hyperactivity Disorder (ADHD) and conduct disorder, Alzheimer’s disease, anemia, anxiety disorder, arthritis, atrial fibrillation, brain injury, chronic kidney disorder , colorectal cancer, chronic obstructive pulmonary disease (COPD), dementia, diabetes, epilepsy, glaucoma, hearing impairment, hyperlipidemia, ischemic heart disease, liver disease (except viral hepatitis), migraine and chronic headache, mobility impairments, osteoporosis, peripheral vascular disease, personality disorders, stroke and transient ischemic attack (TIA), and viral hepatitis.

In order to get more robust Receiver Operation Characteristics (ROC) that reflect the prediction performance also for external datasets, another layer of validation on the test data (\(N = 111,366\)), which was set aside and unseen during the training phase, was applied to derive ROC. With the variable selected via Elastic Net, we developed a final logistic regression model with co-morbidity analysis, and obtained the ROC of the final logistic regression model with co-morbidity analysis on the test data. We then compared the ROC with that of the traditional logistic regression model without co-morbidity analysis. R version 3.1.3\textsuperscript{22} with R software packages, glmnet\textsuperscript{23}, and pROC\textsuperscript{24} were used for this study.
Table 2. Univariate and bivariate statistics of selected demographic, socio-economic, disability registry and co-morbidity variables between the depression case group (N=28,256) and complement comparison group (N=1,085,400). For categorical variables, the observed frequencies of the categories and percentages (twelve-year prevalence) were reported, and for numerical variables, means (and standard deviations) were reported. P-values were of chi-square tests for categorical variables, and t-tests for numerical variables.

<table>
<thead>
<tr>
<th>Demographic and Socio-Economic Variables</th>
<th>Complement Control Group (N=1085400)</th>
<th>Depression Case Group (N=28256)</th>
<th>Total (N=1113656)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>127644 (49.4%)</td>
<td>19238 (68.1%)</td>
<td>54682 (49.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>16 ± 21</td>
<td>18 ± 19</td>
<td>16 ± 21</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Income (Decile)</td>
<td>5 ± 2.5</td>
<td>5 ± 2.5</td>
<td>5 ± 2.5</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Insurance Status - Social Security Beneficiaries</td>
<td>17288 (1.6%)</td>
<td>288 (1.9%)</td>
<td>17576 (1.6%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Residents In Seoul Metropolitan Area</td>
<td>138874 (14.9%)</td>
<td>4138 (14.6%)</td>
<td>133012 (14.9%)</td>
<td>0.275</td>
</tr>
<tr>
<td>Disability Registry Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lumbar Disability</td>
<td>18028 (2.0%)</td>
<td>936 (3.3%)</td>
<td>17964 (2.0%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Neurologic Disability</td>
<td>3528 (0.3%)</td>
<td>258 (0.9%)</td>
<td>3786 (0.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Visual Disability</td>
<td>4321 (0.4%)</td>
<td>197 (0.7%)</td>
<td>4518 (0.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hearing Disability</td>
<td>1467 (0.3%)</td>
<td>772 (0.6%)</td>
<td>2239 (0.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cognitive Disability</td>
<td>2791 (0.3%)</td>
<td>84 (0.2%)</td>
<td>2854 (0.3%)</td>
<td>0.23</td>
</tr>
<tr>
<td>Co-morbidity Variables (Alphabetical Order)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquired Hypothyroidism</td>
<td>24406 (2.3%)</td>
<td>1805 (6.4%)</td>
<td>26211 (2.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Acute Myocardial Infarction</td>
<td>2518 (0.2%)</td>
<td>115 (0.4%)</td>
<td>2633 (0.2%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Alzheimer's Disease</td>
<td>1707 (0.2%)</td>
<td>242 (0.9%)</td>
<td>1949 (0.2%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Anemia</td>
<td>60602 (5.7%)</td>
<td>2875 (10.2%)</td>
<td>63477 (5.8%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Anxiety Disorder</td>
<td>51935 (4.9%)</td>
<td>8980 (31.8%)</td>
<td>60915 (5.6%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Arthritis</td>
<td>213785 (20.0%)</td>
<td>12578 (44.5%)</td>
<td>226363 (20.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Asthma</td>
<td>191650 (17.9%)</td>
<td>6468 (22.9%)</td>
<td>198118 (18.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>3730 (0.3%)</td>
<td>245 (0.9%)</td>
<td>3975 (0.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Attention Deficit Hyperactivity &amp; Conduct Disorder</td>
<td>592 (0.6%)</td>
<td>444 (1.6%)</td>
<td>6367 (0.6%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Benign Prostatic Hyperplasia</td>
<td>3151 (3.1%)</td>
<td>1673 (5.9%)</td>
<td>3332 (3.0%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Brain Injury</td>
<td>15599 (1.4%)</td>
<td>728 (2.6%)</td>
<td>16327 (1.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>3976 (0.3%)</td>
<td>199 (0.7%)</td>
<td>4175 (0.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cataract</td>
<td>66979 (6.5%)</td>
<td>4931 (17.5%)</td>
<td>71910 (6.6%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cerebral Palsy</td>
<td>1004 (0.1%)</td>
<td>18 (0.1%)</td>
<td>1022 (0.1%)</td>
<td>0.12</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>19841 (1.8%)</td>
<td>2324 (8.2%)</td>
<td>22165 (1.8%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic Obstructive Pulmonary Disease (COPD)</td>
<td>34832 (3.0%)</td>
<td>3715 (13.1%)</td>
<td>38547 (3.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic Ulcers</td>
<td>2924 (0.3%)</td>
<td>230 (0.7%)</td>
<td>3154 (0.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>4982 (0.5%)</td>
<td>260 (0.9%)</td>
<td>5242 (0.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dementia</td>
<td>2589 (0.2%)</td>
<td>320 (1.1%)</td>
<td>2909 (0.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>10049 (7.5%)</td>
<td>4708 (16.7%)</td>
<td>87507 (7.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Endometrial Cancer</td>
<td>513 (0.0%)</td>
<td>41 (0.1%)</td>
<td>554 (0.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>8698 (0.9%)</td>
<td>665 (2.4%)</td>
<td>9363 (0.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fibromyalgia and Pain Syndrome</td>
<td>69260 (6.5%)</td>
<td>3995 (14.1%)</td>
<td>73255 (6.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>37216 (3.5%)</td>
<td>2123 (7.5%)</td>
<td>39339 (3.6%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hearing Impairment</td>
<td>40214 (3.8%)</td>
<td>2484 (8.8%)</td>
<td>42698 (3.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>19199 (1.8%)</td>
<td>1408 (5.0%)</td>
<td>20607 (1.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hyperlipidemia</td>
<td>8916 (8.4%)</td>
<td>5700 (20.2%)</td>
<td>94861 (8.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hypertension</td>
<td>57022 (5.3%)</td>
<td>3541 (12.5%)</td>
<td>60563 (5.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>47457 (4.4%)</td>
<td>3692 (13.1%)</td>
<td>51149 (4.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Leukemia And Lymphoma</td>
<td>1461 (0.1%)</td>
<td>64 (0.2%)</td>
<td>1525 (0.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Liver Disease (Except Viral Hepatitis)</td>
<td>127475 (11.9%)</td>
<td>6552 (23.2%)</td>
<td>134027 (12.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>4212 (0.4%)</td>
<td>212 (0.8%)</td>
<td>4424 (0.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Migraine And Chronic Headache</td>
<td>117953 (11.7%)</td>
<td>8192 (29.9%)</td>
<td>126145 (11.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mobility Impairments</td>
<td>11400 (1.1%)</td>
<td>866 (3.1%)</td>
<td>12366 (1.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>13936 (1.3%)</td>
<td>1463 (5.2%)</td>
<td>15400 (1.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pelvic Fractures</td>
<td>3902 (0.4%)</td>
<td>299 (1.1%)</td>
<td>4201 (0.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Peripheral Vascular Disease</td>
<td>22793 (2.1%)</td>
<td>1952 (6.6%)</td>
<td>24745 (2.2%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Personality Disorders</td>
<td>298 (0.1%)</td>
<td>244 (0.9%)</td>
<td>442 (0.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Spinal Cord Injury</td>
<td>819 (0.8%)</td>
<td>590 (2.1%)</td>
<td>8781 (0.8%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stroke And Transient Ischemic Attack (TIA)</td>
<td>49791 (4.7%)</td>
<td>3946 (14.0%)</td>
<td>53737 (4.9%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Viral Hepatitis</td>
<td>43294 (4.1%)</td>
<td>1877 (6.6%)</td>
<td>45171 (4.1%)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Results

The Odds Ratio (OR) plot for the traditional logistic regression model without co-morbidity analysis is presented in Figure 2(a) and the same for the final logistic regression model with co-morbidity analysis in Figure 2(b). It is noticeable that adjusted ORs for the same variables differs between the two models. For example, the adjusted OR of being female is 2.07 from the traditional logistic regression model without co-morbidity analysis, but is 1.63 from the final logistic regression model with co-morbidity analysis. Likewise, the adjusted OR of age is 1.03 from the traditional logistic regression model without co-morbidity analysis, but is 1.01 from the final logistic regression model with co-morbidity analysis. Finally, the adjusted OR of income decile is 1.04 from the traditional logistic regression model without co-morbidity analysis, but is 1.02 from the final logistic regression model with co-morbidity analysis. The ORs for the disability registration variables in the traditional logistic regression model without co-morbidity analysis ranged from 1.03 (cognitive disability) to 1.42 (hearing disability). The ORs for the co-morbidity variables in the final logistic regression model with co-morbidity analysis ranged from 0.78 (acute myocardial infarction) to 5.81 (ADHD and conduct disorder).

Receiver Operating Characteristic (ROC) curve of the traditional logistic regression model without co-morbidity analysis and the final logistic regression model with co-morbidity analysis on the test data, which were unseen during the training phase, are shown in Figure 3. The Area Under the Curve (AUC) of the ROC increased from 0.6992 (the traditional logistic regression model without co-morbidity analysis) to 0.7818 (the final logistic regression model with co-morbidity analysis). Selected performance measures for the 30-predictor co-morbidity model, including sensitivities, specificities, Positive Prediction Values, Negative Prediction Values, Accuracies, and F measures for nine distinct threshold points on the ROC are shown in Table 3.

Figure 1. The plot obtained from cross-validation of Elastic Net, showing the change of the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) with different λ (in log scale) with α of 0.75. The numeric values above the plot indicates the number of variables selected in the between 28 and 61. In other words, 28 (when log(λ) is -6.577986) is the minimum number of variables that guarantees the maximum AUC.
Figure 2(a). Odds Ratio (OR) plot of the traditional logistic regression model without co-morbidity analysis. The point values indicate the adjusted Odds Ratios, horizontal lines indicate the 95% confidence intervals, and the asterisks indicate the level of statistical significance (***, p < 0.001, **: p < 0.01, *: p < 0.05). The variables included in both the traditional model without co-morbidity analysis (a) and the final model with co-morbidity analysis (b) are highlighted in yellow. The adjusted Odds Ratios can differ if variable selection is different. For example, the adjusted OR of being female is 2.07 from the traditional logistic regression model without co-morbidity analysis, but is 1.63 from the final logistic regression model with co-morbidity analysis.
Discussion

Given that depression remains significantly under-diagnosed in all settings of the healthcare system, data-driven prediction models can play an important role in the screening of depression patients. Although previous work has shown promising results on the ability to predict future diagnoses of depression, such models have not explicitly applied co-morbid medical conditions as independent predictors. Depression is a characteristic disease which can be affected by many medical conditions, and can also affect certain medical conditions. In 2013, psychological factors affecting medical conditions (PFAOMC) was included as a new diagnosis in DSM-V. PFAOMC are the factors which may precipitate or exacerbate the medical condition, interfere with treatment, or contribute to morbidity and mortality. The mechanism of PFAOMC include promotion of known risk factors (i.e. smoking), influence on the underlying pathophysiology (i.e. bronchospasm in asthma), and the interference on the treatment (i.e. poor compliance). Therefore, addressing the co-morbidities related to depression will be a rationally important step in understanding the course of depression, and the analysis of these co-morbidities will likely improve the performance of depression risk prediction models.

In this study, we showed that the AUC of the ROC increased from 0.6992 (the traditional logistic regression model without co-morbidity analysis) to 0.7818 (the final logistic regression model with co-morbidity analysis), after applying the optimized variable selection from Elastic Net. Because neither questionnaire-based screening results (i.e. Patient Health Questionnaire) nor physician clinical notes are available...
in claims data, there is no direct information about patients’ moods or symptoms. Given this limitation, this improvement could be interpreted very significant improvement, and the inclusion of co-morbidity analysis could be a key component in improving the performance of depression risk prediction models.

Furthermore, since odds ratio estimates change for some variables after adjusting for co-morbid conditions, the adjusted OR in the final logistic regression model with co-morbidity analysis could reflect estimates closer to the truth. For example, the adjusted OR of being female is 2.07 from the traditional logistic regression model without co-morbidity analysis, but is 1.63 from the final logistic regression model with co-morbidity analysis (Figure 2). Given that, females have a higher co-morbidity burden in general, the traditional logistic regression model will give higher OR for females, by not adjusting for co-morbidities.

The one million twelve-year Korea National Health Insurance Service (KNHIS) longitudinal data used in this study has many advantages for analyzing large scale statistical models. As KNHIS is the only health insurance system which covers all Korean citizens, the random sample cohort from KNHIS can be considered as a nationally representative health data. Factors arising from multiple health insurance systems effecting diagnosis of depression (i.e. some health insurance plans might have lower coverage for mental health) can be avoided in the single health insurance system, and therefore higher statistical power can be achieved. Therefore, adjusted ORs from the logistic regression model with co-morbidity analysis may represent the risks of each variable in the population.

Cautions are needed when interpreting the epidemiologic results from this study, however. The large sample size in this study is over-powered to detect small effects, so more emphasis should be placed on the magnitude of estimates rather than the statistical significance. Furthermore, the operational definition of depression case group is based on the diagnosis codes in the claims data. Therefore, the depression risk prediction model in this study is predicting the probabilities of each person’s visiting physicians and diagnosed as depressed by physicians, and this will limit the ability of detecting the underdiagnosed depressed population. However, it is noticeable that the findings are consistent with previous studies revealed the relationship between co-morbidities and depression in Korean population with cross-sectional survey study, as well as Korean Longitudinal Study of Aging.

In order to develop a better depression risk prediction model which can also address the currently underdiagnosed depressed population, reaching out to the underdiagnosed depressed population with gold standard screening tools will be necessary. Further work is also needed to investigate possible difference in the co-morbidity patterns in different gender, age-group, and socio-economic status. Higher prevalence of depression among female has been discussed to be related to both biological and environmental factors. Features of depressions can also vary among different age-groups, and certain age-groups may have additional risks. Socio-economic factors of depression and disparity in depression treatment are also very important topic in public health.

Additional research is needed for optimizing the chronic conditions clusters, or categories. Although CMS-CCW algorithm is a well validated algorithm using ICD codes, optimized clusters developed using insurance claims data might be different when compared to actual clinical manifestation of depression. Even within the clinical practice, the disease classification or categorization can differ among various clinical specialties and subspecialties. Therefore, optimization for co-morbid conditions clusters will be needed for better prediction models. Furthermore, integration of medication prescription data will allow better operational definitions with lesser false positives. Further research is also needed for variable interactions (i.e. epilepsy of young female may have different effect from epilepsy of elderly male), as well as time-to-event analysis (i.e. Cox Proportional Hazard regression), dealing with time-dependent covariates.

Although the focus of this study was on the prediction of the existence of depression based on health insurance claims data, further studies will be needed to confirm if co-morbidity analysis can also improve the performance of the prediction model for treatment response, or prediction model based on lexical data, as well as on electronic health records.

**Conclusion**

In conclusion, the inclusion of co-morbidity analysis could improve the performance of risk prediction for depression, and the co-morbidity adjusted ORs may indicate the true independent OR of each predictor variable. Further studies will be needed to cover the currently underdiagnosed depressed population, as well as optimizing the chronic conditions clusters.
Acknowledgement

We thank Kyuwoong Kim and Jooyoung Chang at Seoul National University College of Medicine for assistance with data management and collaboration for this research. The data used in this study were provided from Korea National Health Insurance Service - National Sample Cohort (NHIS-NSC) 2002–2013. This study was supported in part by funding from NIH R01MH105384, AHRQ R01HS020377, and UL1 TR000457-06.

Correspondence

Sang Min Park <smpark.snuh@gmail.com> and Jyotishman Pathak <jyp2001@med.cornell.edu>

References

Parental Perceptions, Experiences, and Desires of Music Therapy
Ha-Kyung Kong, Karrie Karahalios, PhD
University of Illinois Urbana Champaign, Champaign, IL, United States

Abstract

Music therapy (MT) is a therapeutic practice where a therapist uses music to enhance the life quality for their patients. Children have an innate enjoyment of music, making music an effective medium for exploring their potential. In this study, we explore the parental perception of MT through an online survey. Contrary to the public perception that MT only addresses emotional needs, 47 out of 59 parents reported seeing improvements in other areas including behavioral, cognitive, linguistic, and social changes. All but one parent indicated that they would recommend MT to others. The survey results further revealed that even parents of children participating in MT had misconceptions regarding MT, which we describe in the paper. Parents reported inaccessibility and cost as other major limitations surrounding MT adoption. We conclude by discussing how technology solutions could mitigate issues with definition, distance, and cost, while maintaining the benefits of MT.

Introduction

Music Therapy is defined by the American Music Therapy Association (AMTA) as “the clinical and evidence-based use of music interventions to accomplish individualized goals within a therapeutic relationship by a credentialed professional who has completed an approved music therapy program.” This definition provides three qualities of MT that distinguish it from therapeutic music, a broader use of music for alleviating a physical, emotional, or mental concern. First, MT is evidence-based and has a strong research foundation. Secondly, the interventions are individualized with a therapeutic goal and conducted over several planned sessions. Thirdly, it is practiced by a Board-Certified Music Therapist (MT-BC), and the therapy occurs within a therapist-patient relationship.

Although MT has been established as a profession for over 60 years, there is “a lack of recognition and understanding of what music therapy is and its benefits [...] many people still have not heard of the profession.” Even people who have heard about MT hold widespread misconceptions about MT including: the client must be musically inclined, music therapists are not real therapists and cannot handle “serious issues,” and more. In actuality, MT can benefit a wide spectrum of people ranging from premature infants, children with autism, adults with traumatic injuries, to older adults with Parkinson's disease. MT is especially effective for children since they naturally enjoy and respond to music; this innate and universal musicality has been termed as “The Music Child” by Nordoff and Robins, and has been used to explore “receptive, cognitive, expressive, and communicative capabilities” of children with developmental delays. Many parents avidly testify about the results they have seen, and research studies show the effectiveness of MT for children and adults. This discrepancy between the potential benefits of MT and the lack of public awareness calls for an exploration of the landscape of MT and development of ways to overcome the challenges that lead to the underuse of MT.

Our research goal is to explore the parental experience of MT through an online survey to identify the current status of MT. Through the research process, we found that even parents of children participating in MT had misconceptions and confusions regarding MT, which we aim to clarify through the paper. We discovered that along with the definition and expectation misconceptions of MT, the inaccessibility and the cost were major limitations surrounding MT-adoption. We conclude the paper by discussing two technology directions that might address the definition, expectation, and accessibility challenges.

Music Experiences and Origins of Approaches in MT

Music therapists use four fundamental music experiences in MT: improvisatory, re-creative, composition, and listening. A MT approach can use one or more music experiences to meet the clinical goals of the session. Darrow, a professor in MT, categorizes the approaches based on their origins are in music education, psychotherapy, or medicine. In this section, we provide a description of each music experience with a list of possible clinical goals for each experience. The clinical goals and the intervention used vary greatly depending on the client's diagnosis and condition, so the list of clinical goals and approaches presented here are not exhaustive.

1) Improvisatory MT focuses on the spontaneous creation of music such as song improvisation and body improvisation. Clinical goals of improvisatory MT experiences include developing creativity and group skills, and establishing non-verbal channels of communication and self-expression.
2) **Re-creative** MT involves singing and playing pre-composed music and is used to work with clients of wide ranging diagnosis from aphasia, attention deficit hyperactivity disorder (ADHD), and cerebral palsy. Clinical goals of re-creative MT experiences include increasing self-confidence and attention span, and developing a sense of mastery.

3) **Composition** methods involve creating vocal and instrumental pieces as a means of self-expression. Clinical goals of composition include raising self-esteem through decision making and reaffirmation of the therapist, “development of self-image, and dealing with emotional loss and trauma”.

4) **Listening** MT is a receptive musical experience involving listening and responding to live or recorded music. Clinical goals of listening MT include evoking and supporting emotional responses, promoting self-expression, and verbal processing of meaningful ideas.

These music experiences are used in different approaches in MT that come from three different origins: a) music education, b) psychotherapy, and c) medicine. We provide below an approach from each of the origins and how the four music experiences are used within each approach.

a) **Orff-Based MT** approach is based on Orff Schulwerk, a multisensory music education approach based on learning by doing that emphasizes using elemental music in which everyone is able to participate. In an Orff-based MT session, the children start by exploring the musical possibilities, and then are led to imitate a pattern the music therapist presents (re-creative). Once they successfully imitate, they improvise using body percussion (eg. snapping), non-pitched percussion (eg. drums), then pitched percussion (eg. xylophones) and the music therapist organizes the musical ideas to fit the goal of the session (improvisational). Children learn turn taking, looking out for subtle cues, and collaborating towards a common goal through this exercise. The final stage is creation where they combine what they have learned in the previous stages (composition). The choice making process and affirmative response from the music therapist raises the children's confidence and can improve their decision-making skills.

b) **Bonny Method of Guided Imagery and Music (BMGIM)** originates from psychotherapy. Since this approach involves verbal reflections of the clients' experiences, clients have to be cognitively and physically capable of verbalizing their thoughts. A BMGIM session consists of a preliminary conversation (to build rapport and to set the tone of the session), relaxation, music listening, and a post-session review (listening). The listening section begins when the music therapist turns on pre-recorded music for the child that is pre-selected based on its therapeutic usefulness. The music therapist provides an opening imagery scene and lets the child verbalize the imagined scenes throughout the music while providing encouragement and empathy. The child is encouraged to make interpretations and parallels to his/her life in the post-session review. Through the focused and pre-planned discussion, music promotes psychodynamic healing and growth of interpersonal relationship between the therapist and the child.

These types of psychotherapeutic music listening techniques have to be clearly distinguished from informal therapeutic experiences involving music (ex. listening to relaxing music at home to lighten one's mood) and formal therapeutic music (ex. bedside musicians playing for patients in the hospital). Sound therapies, where the inherent properties of music itself are considered as the agent of change, are not considered MT by music therapists either. However, some branches of sound therapy refer to themselves as MT. In fact, our study participants termed sound therapies as MT. The schism between music therapists' view of clinical MT and the outsiders' view of MT is elaborated further in the Discussion Section.

c) **Neurologic Music Therapy (NMT)** is a medical approach to MT developed by Dr. Michael Thaut and his colleagues that investigates the effects of music on the brain and behavioral functions. NMT methods are used for sensorimotor training, cognitive training, and speech and language training through different properties of music and sound such as pitch, timbre, and rhythm. Melodic intonation therapy (MIT) is a type of NMT that has been effective for treatment of children with apraxia of speech. The learning occurs in a gradual pattern while music therapist taps the rhythm. The phrase is first hummed by the music therapist and imitated by the child through signing, then repeated “speech singing,” and eventually spoken in normal speech (re-creative).

Based on this wide range of music experiences and approaches in MT, different clients have different perceptions, experiences, and desires of MT. In this study, we surveyed parents whose children had participated in MT to address the following research questions:

**RQ 1** Why do parents choose music therapy for their children?
RQ 2 Why do parents stop music therapy?
RQ 3 What benefits do parents expect from music therapy sessions?
RQ 4 What benefits did parents see through music therapy sessions?
RQ 5 What changes do parents want to see in the field of music therapy?

Methodology

Our study was conducted through an anonymous online survey on parental satisfaction of MT in the U.S. and South Korea. The survey was designed based on existing MT literature with input from music therapists and consisted of 19 closed-ended questions and 8 open-ended questions that explored the types of MT in which the child had participated as well as the parents' expectations and thoughts on MT. We recruited participants through mailing lists; posts on online forums (e.g., Reddit); flyers in cafes, libraries, and healthcare centers around three large universities; and music therapists. Additionally, we contacted approximately thirty-five health care centers offering music therapy in the U.S. and South Korea through email. Recruiting participants proved difficult and took over a year. We first started by recruiting participants in the U.S. and expanded our study to include participants from South Korea when Korean music therapists showed interest in our study. The Korean parents were mainly recruited through music therapists and therapy centers in South Korea. As an incentive, each participant was entered into a raffle for a $50 Amazon gift card.

Participants

Between March 4, 2014 and February 6, 2015, we received 59 responses through a Korean and an English version of an anonymous online survey: 38 from South Korea and 21 from the U.S. The details of the survey results are provided in Table 1. All but two (n=53) of the respondents who indicated their relationship to the child who participated in MT were mothers; one respondent was the father and one respondent was a legal guardian. The majority of the parents surveyed (n=46) had one or two children currently living at home, with the number of children at home ranging from one to seven (Eng: $\mu=2.19$, $\sigma=1.47$; Kor: $\mu=1.58$, $\sigma=0.80$). Ages of the focus children ranged from 15 months to over 25 years old, with the mean age of 12.

Procedure

Survey Design

After presenting the consent form on the introductory page, the survey collected demographic and basic information about the focus children. Next, the survey asked about the respondent's relationship to the child who had taken MT and the total number of children at home to probe the possible influence on the amount of parental involvement in MT. The survey then targeted RQ1 by asking where the participants had heard about MT, what prompted them to try MT, and what other therapies they had tried. To answer the second research question, the survey asked if the child is currently receiving MT services. If no, the survey asked for the main reasons for discontinuing MT.

The next seven questions focused on the specifics of the MT sessions by asking how long the child has received MT, the type of music experiences in MT sessions, and the location, frequency, duration, and the type of MT sessions (personal, group, family, other). The survey then asked the participants to rate the parental involvement during and outside of MT sessions, the music therapist's understanding of the child's needs and strength, and whether the child looked forward to the MT sessions on a five point Likert scale.

Next, the survey presented multiple-choice questions to ask for areas they would most like to see improvements through MT (for RQ3) and for areas they saw a satisfactory improvement (for RQ4). We followed these questions open-ended questions to elaborate on the improvements they saw and how they would improve MT. We then asked whether they would recommend MT to family and friends. The survey ended by asking for the participant's birth date to filter out spam and ensure that the participant was over 18. We also filtered out responses that did not answer any of the open-ended questions.

Data Analysis

We analyzed the open-ended questions by establishing a list of common themes through open coding and categorizing each response into one of the themes. The emerged themes are mentioned in the results section. Two raters categorized all 90 responses independently; the intercoder reliability between the two coders was substantial,

1 Parental involvement outside of MT sessions includes, but is not limited to, learning new ways of using music at home and eliciting skills that the child learned during MT sessions.
κ = .77, p < 0.05. They resolved the differences by discussing each response till they came to an agreement for all the responses. All of the Korean open-ended questions were translated by one of the researchers and proofread by a graduate student who has an English Education degree from a Korean university.

We analyzed responses from the two surveys (English and Korean) separately and did not make comparisons between the two samples because the participants were not randomly sampled from the two populations. Four of the five Likert scale responses were combined into a composite score given that the results from the two samples are significantly equivalent based on the two one-sided tests for equivalence (p < 0.05, ε = 1). The last Likert scale response is reported separately. The responses from the two countries’ surveys were also combined to compare the desired areas of change to actual areas of change. We used a paired-sample test that compared each respondent’s desired changes and actual changes. Fourteen Korean responses relating to the music therapists were discarded. This is because they were collected directly through a music therapist and it is possible that those respondents felt pressured to answer more positively to questions regarding that music therapist. That left us with 21 English responses and 24 Korean responses for those two questions relating to the music therapist.

Results

The average response rate for each question in the survey was very high (97%) with all but one question getting a response rate higher than 95%; the last opened ended question asking for areas for improvement had a 58% response rate. Most of the parents first learned of MT through a health care provider (e.g., doctor, therapist), friends, and online sources (e.g., blog, SNS, search engine). Other sources of information came from school, family, rehabilitation centers, magazine articles, and television. The location of the sessions was split between home (n=8; 38%) and clinics (n=7; 33%) for the English survey while the majority of the Korean survey respondents (n=34; 89%) answered clinics. Personal therapy was the most common form of therapy followed by group therapy for both the American and Korean clients. One respondent answered that the child participated in both personal and group sessions. However, one parent commented that they would prefer if her child had access to a group MT session, indicating that not everyone has access to all types of MT.

    | Age of the child (in years) | Sessions / Week |
    |-----------------------------|-----------------|
    | 0-5            | 6-10           | 11-15 | 16-20 | 20+   | 1   | 2   | 3   | 4 or more |
    | 5 | 5 | 6 | 13 | 3 | 10 | 4 | 6 | 3 | 4 | 17 | 27 | 1 | 9 | 1 | 1 | 2 | 1 |
    | Total number of children at home | Location of MT |
    | 1 | 2 | 3 | 4 | 4+ | school | clinics | home | other |
    | 8 | 14 | 7 | 17 | 3 | 4 | 2 | 0 | 1 | 0 | 3 | 0 | 7 | 34 | 8 | 3 | 4 | 1 |
    | Duration of MT (in months) | Type of MT |
    | <= 12 | 13-24 | 25-36 | 37-48 | 48+   | personal | group | family | other |
    | 9 | 19 | 3 | 4 | 4 | 5 | 1 | 4 | 2 | 5 | 16 | 31 | 4 | 7 | 0 | 0 | 1 | 0 |
    | Became aware of MT through | Minutes / Week |
    | health care provider | family | your child’s school | friend | online | less than 30 | 30 to 60 | 60 to 90 | more than 90 |
    | 6 | 18 | 2 | 1 | 2 | 1 | 3 | 5 | 2 | 5 | 3 | 0 | 14 | 25 | 1 | 9 | 3 | 4 |
    | Other therapies taken | Reasons for discontinuing MT |
    | occupational therapy | speech therapy | behavioral therapy | none | other | lack of improvement | schedule conflict | reached max benefit | availability of MT | cost of service |
    | 19 | 3 | 17 | 23 | 8 | 10 | 1 | 5 | 6 | 17 | 1 | 0 | 2 | 0 | 2 | 0 | 4 | 0 | 1 | 1 |
    | Child still participating in MT | Willing to recommend MT to others | Music experiences |
    | yes | no | yes | no | improvisatory | re-creative | composition | listening | other |
    | 11 | 35 | 10 | 3 | 19 | 38 | 1 | 0 | 12 | 35 | 15 | 30 | 3 | 14 | 14 | 23 | 1 | 12 |

Table 1. Table of survey results. Responses reported as Eng | Kor (out of 21 | 38)
Most of the children received MT once a week or twice a week. The duration of the sessions was primarily 30 minutes to 60 minutes (n=39) or 60 minutes to 90 minutes (n=10). This frequency and duration of MT is common for developmental treatments while rehabilitative MT can be more intensive. For example, the regimen for MIT, previously mentioned in section 2 can be as intensive as 1.5 hrs/day for five days/week as it utilizes brain plasticity for recovering speech”.

In addition to MT, parents had also tried speech therapy (n=40), occupational therapy (n=21), and behavioral therapy (n=18). Eight of the Korean parents had tried art therapy. This can be explained by the options of therapies commonly offered in social welfare centers and clinics, the location of MT sessions for the majority of the Korean clients.

1. Motivations for starting and discontinuing music therapy
Motivations for starting MT mostly consisted of the diagnosis of the child. Six parents specified autism spectrum disorder (ASD); eight parents specified speech delay, disorder, or a non-verbal child; six specified developmental delay as the diagnosis. Other parents reported Down Syndrome (n=2), ADHD, episodes of seizures, anxiety disorders, Williams Syndrome, and encephalopathy as reasons for starting MT. Parents further reported seeking MT to see specific benefits such as reduced frequency of seizures, improved self-expression, and stress relief. Six parents indicated the children's interest in music. Respondent 15 wrote, “Due to seizures music was one of two ways she would respond to stimuli.” Similarly, Respondent 18 wrote, “my daughter has Down syndrome, has many areas of delay, and also loves and responds to music.” This response to music regardless of the child's delay in other areas is a key factor in “the Music Child” and how MT can bring out the child's potential.

Out of 13 parents who reported that their children were not currently receiving MT services, four reported unavailability of MT, three reported a conflict in schedule, two reported that they had received maximum benefit. The others reported unwillingness of the child to participate, ineffectiveness of therapy, cost of service, and recommendations from the child's teacher to stop. Throughout the survey and other correspondence with parents of children who had participated in MT, it became apparent that one of the main limitations of MT was the unavailability of MT as 8 participants listed it as a change they would like to see in MT, and 7 out of 13 who discontinued MT had difficulty accessing MT.

2. Experiences in music therapy
The most common types of music experience were improvisation (n=47) and re-creative (n=45). Listening was also widely used for the American patients (n=37) while composition was not as commonly used in either population (n=17). Most parents (n= 47) reported that the child had experienced more than one type of musical experience in MT. Parents were satisfied with the level of their involvement during and outside of MT sessions overall. The mean score to the statement “I was satisfied with the amount of parental involvement during the session” on a Likert scale from 1(strongly disagree) to 5 (strongly agree) was 3.93 (σ= 0.95), and the mean score for involvement outside of the session was 4 (σ= 0.84). The scores were significantly greater than a neutral response of 3 according to the Wilcoxon signed rank test (Inside session: V = 579.5, p-value < 0.001; Outside of session: V = 601, p-value < 0.001). Parents thought that the music therapists' understanding of the child's strengths was satisfactory (µ= 4.36, σ= 0.86) and that the music therapists' understanding of the child's needs was satisfactory (µ= 4.47, σ= 0.81). Both groups reported that the children looked forward to the MT sessions, but the average value was slightly higher for the Korean respondents (Eng: µ= 3.76, σ= 1.22; Kor: µ= 4.29, σ= 1.12).

3. Desired changes and satisfaction
To address RQ3 and RQ4, we asked the participants to indicate the areas of desired changes prior to MT and perceived outcomes of MT, respectively. The results for these questions, “In which of the following areas would you most like to see improvements through music therapy? Select up to 3 answers” and “In which of the following areas did you see a satisfactory improvements through music therapy? (Select all that are true),” are shown in Figure 1.

The seven developmental areas used to indicate desired and perceived changes were: cognitive, behavioral, language, physical, social, emotional, and musical. These areas were chosen based on the target areas listed by AMTA and prior research on MT outcomes. Emotional, behavioral, and cognitive were respectively the first, second, and third areas in which Korean parents would like to see changes, while behavioral and cognitive were first and second in the English survey. Language and emotional tied as the third. One hypothesis explaining the Korean parents' emphasis on emotional improvements is that the intense academic pressure in South Korea might adversely affect children's emotional stability.
Figure 1. Y-axis is the percentage of parents who indicated the area for desired (left) or actual (right) changes. If all parents marked one area, its total value would be 2.0 (1.0 for Korea and 1.0 for the U.S.). Note that parents could mark all areas of actual changes, but only the top three desired changes.

If a parent selected a desired area for change and selected that area again as a perceived improvement, we noted that that area was met with satisfaction via MT. Desires for musical changes were most likely to be satisfied, with 89% of parents who checked it as desired also checking it as a perceived outcome. Desires for emotional changes were the second most likely to be met at 79%. Desires for cognitive improvement were least likely to be met through MT with 44% of parents wanting to see cognitive changes reporting perceived changes. In one noteworthy case, a parent reported continued satisfaction for their 22-year old child with autism, who had participated in MT for 15 years (Respondent 21). In addition to the three standard therapies mentioned in our survey – occupational therapy, speech therapy, and behavioral therapy, the child had participated in numerous other therapies including “physical therapy, interactive metronome therapy, cranio-sacral therapy, hippotherapy, pet (dog) therapy, Fast ForWord, Listening Program, [and] auditory integration therapy.” The parent had checked cognitive, behavioral, and emotional for the desired changes and all the categories for perceived outcomes showing high satisfaction with MT.

All but two of the participants responded that they would recommend MT to family and friends (n=57) indicating their overall satisfaction with MT. One participant stated that she would not recommend MT because she perceived no changes, and one participant did not provide a response.

4. Benefits of music therapy

The survey asked, “Could you give an example of an improvement? (ex. my child can focus longer)” to investigate specific changes parents perceived through MT. Out of 56 parents who provided a response, many reported changes in multiple developmental areas. The majority of the reported changes were emotional changes (n=19). Parents reported emotional stability (n=8) and enhanced emotional expressiveness (n=7). For example, Respondent 12 wrote, “We received receptive & expressive emotions, listening, and increased happiness and a connection via eye contact/smiles.”

Seventeen parents reported behavioral changes including an increase in the following of directions and decreased frequency of tantrums. Respondent 33 wrote that her child “doesn’t throw tantrums as often, doesn’t hit the younger sibling as often, does homework, washes more often, [and] started going on school field trips (translated from Korean).” Similarly, Respondent 37 wrote, “ADHD symptoms became less severe, reduced impulsive and violent behaviors and shouting. Reduced roaming around, throwing fits, and gluttony (translated from Korean).”

Language changes were also prevalent in the responses (n=14). Parents reported improvements in vocalization, timing and rhythm of speech, pronunciation, and self-expression. Respondent 6 stated that “[her] son could speak in sentences instead of phrases. Better attention. Quicker processing of questions he was asked and shorter time to respond.” Relating to the phenomenon that some people who struggle to speak can still sing the words, Respondent 17 wrote, “[my daughter] will sing to talk to us, before she couldn’t tell us her needs.” This activation of language through singing is often cited as a major strength of MIT mentioned in Section 2. Other major reported improvements included cognitive (n=12) and musical (n=10) improvements such as learning how to read music. Respondent 34 wrote that her child “sang for the first time, unique accents and pronunciation became better, more focused when studying, doesn’t get irritated or angry as often (translated from Korean).”

5. Desired changes in the field of MT

Parents reported various limitations surrounding MT in response to the final question, “How can music therapy be improved for your child?” Of the 34 submitted responses, the following themes emerged: more
individualized/diverse programs (n=9), accessibility (n=8), awareness (n=6), lowered cost (n=4), and acknowledgement from organizations (n=3).

Parents often sought out specific programs, environments, and opportunities that addressed their children's needs such as receiving therapy in a larger, sound-proof space. Respondent 21 requested more social and language-related programs; Respondent 33 requested opportunities for her child to play ukulele with other children. Two parents requested programs that are currently available in other regions but not in their own (ex. group therapy). Some parents currently did not have access to MT at all. Concerning the accessibility of MT, one mother wrote, “We need providers in central Illinois that [are] affordable and available. We found no one in the area to help. We travelled to Texas for our initial treatment and then continued at home with very little outside help. I am not a music teacher and this took the fun element out of the therapy for my son.” Other parents wanted to raise the awareness of MT as the treatment had been effective for their children.

Cost is a major concern of many parents. One respondent (8) stated, “We paid $100 for a consultation with an occupational therapist. She played special music for my daughter during the session them [sic] recommended that I purchase a $25 mp3 soundtrack, $50 headphones and a $150 MP3 player because the soundtracks were only sold as mp3. I'm interested in music therapy but not at that price!” This burden of cost could be alleviated if the service was run by schools. However, schools are often unfamiliar with MT or skeptical of its effectiveness. Respondent 6 stated that one way to improve MT is through “convincing my child[sic] school administration on how beneficial it is. Having them also realize it's a related service recognized by the Department of Education.” Respondent 10 mentioned that “It would be nice if schools would recognize music therapy as an educational tool for children with disabilities. As it stands school districts do not want to allow it in schools or to be used within a school environment.” Respondent 32 stated that the reason for discontinuing MT was because the child’s kindergarten teacher told them to stop. No further explanation was given to explain this recommendation.

Two respondents indicated that there was nothing they would change. Respondent 14 wrote:

Our therapist is amazing. She understood our daughter's strengths and used those to help her speak. [...] For example “I want to go outside” is to the tune of Mary Had a Little Lamb. Because it is music, our daughter will now sing that when she wants to go outside to play instead of just fussing. She now sings when she is hungry, thirsty, wants to play the piano, wants TV, or to listen to music, or needs a diaper change.

Discussion

The results of this anonymous online survey convey parents’ original desires and experience with MT and suggestions for the field. Additionally, the results and the process of research revealed a sense of confusion and unawareness of the exact definition of clinical MT. The schism in the MT and sound therapy communities aggravated the confusion of parents and schools. In the following section, we shed some light on this confusion. Then, we propose how technology could enhance the field of MT based on the results from the survey and our discussions of this survey with music therapists.

Ambiguities of music in health and healing

Our survey responses highlighted that parents of children who received Samonas Sound Therapy classified it as MT although it does not meet the third criteria established above. Samonas Sound Therapy is an auditory intervention program that uses physical properties of processed music and sound “to re-map and restore the brain's ability to process sound.” Thus, specific music features such as shifted pitch are the agent of change rather than the musical experience as in clinical MT. Since the healing power is inherent in the music in sound therapy, a child can listen to a selected recording at home without a therapist-patient relationship and still receive therapeutic benefits of Samonas Sound Therapy. Furthermore, although sound therapy often requires training as well, the certification processes are considerably different; for example, the authors of this paper received “Therapeutic Listening” certification via a 12-hour online training course. Most of the participants in the course were trained occupational therapists who were interested in extending their programs to include sound therapy. To become a Music Therapist-Board Certified (MT-BC), one must obtain a bachelor's degree or higher in MT, perform 1200 hours of clinical training, and pass the national board certification.

Music therapists with whom we corresponded confirmed that sound therapy was absolutely distinct from clinical MT. They considered it important to distinguish therapeutic music and sound therapy from MT because this mislabeling could diminish the credibility of MT. Their main concern was the lack of evidence-based research and the lack of therapist-patient interaction in sound therapies. Thus, although sound therapy is also an authentic way of
using music for healing, it should be distinguished from MT because the two disciplines’ fundamental philosophies in the use of music are different.

**Assistive Technology for MT**

The value and the need for MT at a distance became apparent after the parental reports in this study aligned with comments we had received in our lab over the year from parents missing appointments or struggling to reach our speech and hearing clinic. In this study, we found that the unavailability of MT was the leading cause for discontinuing MT with four people out of 13 who discontinued MT articulating this in the survey and eight parents listing accessibility as the area of improvement in an open-ended question. In 2011, there were 3,352 AMTA members in the U.S. As the majority of them work in urban regions, people in rural areas had limited access to MT. To date, few teletherapy approaches to MT have been explored. This idea of using technology to augment MT, however, was met with a strong opposition from select music therapists. The main objection was that technology, or rather, a lack of direct therapist-patient involvement, might negatively impact the outcome of a therapy session. In this section, we explore two technology directions for MT that address the limitations surrounding MT while attempting to maintain the benefits of MT. These are (1) teletherapy and (2) clear and definitive MT resources that raise awareness, advance understanding of MT, and enhance therapist-parent communication to encourage continuous therapy at home.

**Teletherapy: supporting remote MT**

Teletherapy, also referred to as e-therapy, is emerging as an alternative that offers equal access to different therapies regardless of one’s geographical region. Although teletherapy can be synchronous or asynchronous and comes in different modalities (e.g., email, chat, and video-conferencing), in this paper, the term “teletherapy” refers to the use of videoteleconferencing (VTC) to deliver therapy sessions when distance separates the client and the therapist. Clinicians have been investigating telecommunication methods for patients in remote areas since the mid-1950s. Previous teletherapy studies have shown the acceptability and feasibility of teletherapy for children and youth. One such study measured Therapeutic Alliance (TA), the relationship between a healthcare professional and a client in teletherapeutic settings. The maintenance of TA is especially important in MT since the therapeutic changes are made through the relationship between the therapist and the client. The results supported that “TA can be developed in psychotherapy by videoconference, with clients rating bond and presence at least equally as strongly as in-person settings across a range of diagnostic groups.” This shows that MT could potentially be conducted at a distance while maintaining the essential therapist-patient relationship. One possible MT teletherapy scenario: MIT can be conducted using VTC where the therapist leads a singing session while the child sings along wearing an augmented wristband or a glove that taps the beats on the wrist. This allows the two research-supported therapeutic elements of MIT – melodic phrasing and rhythmic tapping – to be incorporated in teletherapy. Bandwidth and network limitations in rural areas pose a challenge for this approach. Remote music collaborations are difficult as sound lags disrupt synchronization, which in turn may erode the benefits of MT. These limitations suggest more abstracted low bandwidth interactions between the remote participants that do not rely on traditional VTC.

An alternative and perhaps preferred usage of technology for addressing the inaccessibility of MT is educating the parents in real-time so that the therapy could be conducted at home under the supervision of the therapist, thus minimizing travel. The idea is presented in “Coaching Parents of Young Children with Autism in Rural Areas Using Internet-Based Technologies: A Pilot Program” and could be extended to explore how Internet-based technologies could enhance communication between the music therapist and the parent to encourage continuous treatment. In this scenario, a clinician observes the parent-child interaction via VTC and gives real-time parental guidance via an earpiece worn by the parent to assist the parent in a parent-child clinician guided session. This approach mitigates the remote syncing challenges of the remote participation approaches while encouraging direct social interaction between the parent and child.

**Information organization: raising awareness and receiving acknowledgement**

Out of 34 parental suggestions for MT, nine (26%) addressed the need to promote MT through raising awareness for the general public (n=6) and receiving acknowledgement of MT efficacy from schools (n=3). Even parents themselves experienced difficulty in understanding the exact definition of MT – shown through parents who mistook sound therapy and MT – and the interventions used in MT – shown through two parents who responded that they did not know what methods were used during their children’s therapy sessions. The music therapist who sent us a collection of responses further indicated that some parents do not understand the precise purpose of MT and rather perceive it as a method for relieving a child’s stress. This limited view of MT can be problematic as parents and the
general population undermine the effects of MT and do not consider it for treatment in non-emotional areas. Even if a person is interested in MT, it is currently difficult to access a list of the major methods and interventions used in MT and how each method of MT benefits the clients. The AMTA website offers general information on how MT can be applied to a specific population or situation (e.g., young people, Alzheimer's patients, pain management). However, it does not provide specific interventions that are frequently used for young children (e.g., Nordoff-Robbins Music Therapy). Although relevant information can be found by searching online and reading academic journals, the search is not only burdensome but also confusing for time-challenged parents who are not familiar with the area and lack a starting framework.

We propose a site where a collection of interventions categorized by music experience, origins, target population, clinical goals, etc. is presented with case studies and research findings. This MT archive would help clarify the clinical definition of MT to the public, and would be useful for informing schools and other organizations of the uses and benefits of MT. It could also have evidence based MT research for those who are interested in pursuing the literature to date. After logging in, parents could also view a summary of their children’s MT sessions and a list of parent-child exercises that can be done at home to enhance their understanding of MT sessions and enable continuous treatment. The official Autism Speaks site is an exemplary site that presents everything starting from basic information such as the definition and symptoms of autism to resources for parents and novel research findings.

Since the site has many of the elements we are exploring, it could be used as a model for raising the public awareness and receiving organizations' acknowledgement.

Limitations
While the findings of this study contribute to the understanding of parental experience of MT and prompt a discussion on technological applications, this study is not without limitations. The recruitment of the participants through music therapists and clinics might have biased the results to over-represent certain approaches of MT over others and participants predisposed to the benefits of MT resulting in overly positive responses. Even with aggressive advertising for approximately one year, we struggled to reach participants and did not have enough participants from each country to make a cultural comparison. The non-random sampling, the small sample size for each population, and the self-selection bias might limit the generalizability of the findings. However, we found commonly recurring themes in responses of participants from both countries suggesting consensus for the specific benefits and limitations of MT.

Conclusions
In this paper, we presented findings from a survey designed to make sense of parents’ expectations and experiences of MT for their children. The study contributes to the understanding of MT in three areas: survey findings highlighting satisfaction, limitations, and desires of MT; the discovery of the parent-clinician misalignment of the meaning of music therapy; and a discussion of future directions using simple existing technologies to create a united MT face and to create computer-mediated technologies to explore remote accessibility for MT.

Through the survey we found that parents: (1) expressed strong satisfaction with MT and 57 out of 58 parents stated that they would recommend MT to others; (2) indicated that they would like MT to be more accessible, affordable, and publicly recognized. Throughout our survey preparation and analysis, we discovered parents and MT clinicians had different models representing music therapy – parents often included therapeutic music techniques in their MT umbrella. Finally, we presented two future directions to address the MT limitations that arose from the survey and research process: teletherapy for making MT more accessible and information organization for raising the awareness and being acknowledged by the public and institutions.

References
Ensembles of NLP Tools for Data Element Extraction from Clinical Notes

Tsung-Ting Kuo, PhD¹, Pallavi Rao, MS², Cleo Maehara, MD, MSc³, Son Doan, PhD¹, Juan D. Chaparro, MD¹, Michele E. Day, PhD¹, Claudiu Farcas, PhD¹, Lucila Ohno-Machado, MD, PhD¹, and Chun-Nan Hsu, PhD¹

¹University of California San Diego, La Jolla, CA; ²University of California, Davis, CA; ³University of California, Los Angeles, CA

Abstract

Natural Language Processing (NLP) is essential for concept extraction from narrative text in electronic health records (EHR). To extract numerous and diverse concepts, such as data elements (i.e., important concepts related to a certain medical condition), a plausible solution is to combine various NLP tools into an ensemble to improve extraction performance. However, it is unclear to what extent ensembles of popular NLP tools improve the extraction of numerous and diverse concepts. Therefore, we built an NLP ensemble pipeline to synergize the strength of popular NLP tools using seven ensemble methods, and to quantify the improvement in performance achieved by ensembles in the extraction of data elements for three very different cohorts. Evaluation results show that the pipeline can improve the performance of NLP tools, but there is high variability depending on the cohort.

Introduction

Extracting concepts from narrative clinical notes using Natural Language Processing (NLP) techniques is essential to enhance cohort identification processes. Researchers have developed many clinical NLP concept extraction tools (NLP tools), such as cTAKES⁴ (clinical Text Analysis and Knowledge Extraction System) and MetaMap.⁵ An NLP tool may be suitable and powerful for certain concept extraction tasks; there is hardly an NLP tool that is general enough to deal with all extraction tasks. This issue becomes more challenging when the types of concepts to be extracted are numerous and very diverse. For example, in the task of extracting data elements (important concepts related to certain medical conditions, such as encounter information, laboratory tests, imaging tests, and medications) from clinical notes, there are usually several types of data elements that need to be extracted (e.g., 183 data elements in our three cohorts), and different categories of data elements may be better extracted by different NLP tools because of the difference in concept dictionaries or extraction algorithms/pipelines. To address this issue, we propose to apply ensemble methods to integrate NLP tools. Ensemble methods have been proven to be effective to boost the performance of classifiers.⁶ The superiority of ensemble methods has been shown empirically in a wide range of data mining and NLP competitions.⁷,⁸,⁹,¹⁰ Many ensemble methods are available, including basic methods (union and intersection) and advanced ones (Binary Relevance (BR),¹¹ Multi-Label K-Nearest Neighbor (MLKNN),¹² Instance-Based Logistic Regression for Multi-Label (IBLR-ML),¹³ Random k-Labelsets (RAkEL),¹⁴ and Ensemble of Classifier Chains (ECC)¹⁵), spanning a large spectrum of sophistication. Advanced ensemble methods can weigh the importance of component NLP tools by applying machine learning to learn the best set of weights.

Although ensemble methods are empirically known to improve classification performance in several problems, it is not clear to what extent ensembles of NLP tools improve the extraction of numerous and diverse data elements from clinical notes. Several recent studies applied ensemble methods to extract concepts from clinical text.¹⁶,¹⁷,¹⁸ However, most of them focus on the extraction of a few concept types, such as identifying 6 types of medication information,¹⁹ predicting 3 types of concepts (problem, test, or treatment),¹⁰ classifying 9 types of radiology concepts,²¹ or determining 8 concept types related to heart disease.²² It is unclear whether these methods are suitable for dealing with numerous and diverse concepts such as the 183 data elements from our project.

Phenotyping (i.e., characterization of disease states using electronic health records) relies heavily on structured data items, as well as on NLP for data element extraction from narrative text, and is a critical component of precision medicine.²³,²⁴,²⁵,²⁶,²⁷ Our goal was to quantify the improvement in performance achieved by ensembles of popular NLP tools in the phenotyping of three very different cohorts for (1) congestive heart failure (CHF), (2) weight management/obesity (WM/O), and (3) Kawasaki disease (KD). These three conditions are use cases for the pSCANNER (patient-centered SCAlable National Network for Effectiveness Research) clinical data research network, a stakeholder-governed federated network that utilizes a distributed, service-oriented architecture to integrate data from multiple health systems²⁸.
In order to evaluate the ensemble approach, we developed an NLP ensemble pipeline to integrate two popular NLP tools, cTAKES\(^1\) and MetaMap.\(^2\) Our pipeline can integrate these NLP tools using the basic and advanced ensemble methods mentioned previously, and allows us to evaluate the performance of NLP tools when they are applied alone and when they are integrated using different ensemble methods.

**Methods**

**NLP Ensemble Pipeline**

Figure 1 provides an overview of our pipeline. Inputs are clinical notes, which are preprocessed before they are ready for NLP tool extraction of data elements. The output of the NLP tools is integrated by the ensemble methods, and either structured-formatted files or annotated files serve as outputs. The output formats of our pipeline include extraction results such as structured data that are ready to be exported in a Common Data Model (CDM) format (such as the one pSCANNER uses, the Observational Medical Outcomes Partnership CDM (http://omop.org/CDM)), as well as annotation tags in the format used by the Brat Rapid Annotation Tool (BRAT),\(^4\) a Web-based graphical interface for text annotation that allows users to visualize and correct outputs if necessary.

In this paper, we focus on three cohorts of pSCANNER: CHF, WM/O, and KD. Subject matter experts identified important data elements for research on patients with these conditions. To extract these data elements, we integrated two NLP tools: cTAKES\(^1\), an NLP tool for concept extraction from free text clinical notes in an EHR; and MetaMap,\(^2\) a tool for recognizing UMLS\(^9\) concepts in text. These tools cover a wide range of applications of NLP for clinical note concept extraction and clinical note processing.

**Data Elements and Mapping Tables**

We worked with subject matter experts to identify 183 common data elements related to each cohort (50 for CHF, 96 for WM/O, and 37 for KD). The resulting data elements for CHF, WM/O and KD are shown in Tables 1, 2 and 3, respectively. PR and CNH mapped all data elements to the most specific standard codes (SNOMED-CT,\(^6\) LOINC,\(^7\) RxNorm,\(^8\) or UMLS\(^9\)) using the BioPortal web service\(^10\) and created a data-element mapping table for each condition. We used the data-element mapping tables to normalize the output formats of the NLP tools that we integrated. The output standard codes of cTAKES include SNOMED-CT and RxNorm. MetaMap outputs UMLS. These standard codes were mapped to unique data elements in the table. If the output of any tool contained multiple standard codes for an extracted concept, we mapped all standard codes to the unique data elements. This way the outputs of all NLP tools were normalized and ready to be inputs for the ensemble methods.
Table 1. Common data elements for congestive heart failure (CHF)

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms</td>
<td>Congestive Heart Failure</td>
</tr>
<tr>
<td>Encounter Information</td>
<td>First Outpatient Appointment Date, Days Since Symptom Onset, Date Admitted, Date Discharged</td>
</tr>
<tr>
<td>Other Information</td>
<td>Height, Weight, Body Mass Index</td>
</tr>
<tr>
<td>Laboratory Tests</td>
<td>Blood Urea Nitrogen, Brain Natriuretic Peptides, Lipids, Serum Creatinine, Red Blood Cell Count, Serum Albumin, N-Terminal of the Prohormone Brain Natriuretic Peptide, Troponin</td>
</tr>
<tr>
<td>Imaging Tests</td>
<td>Chest X-Ray, Cardiac Magnetic Resonance Imaging, Angio Computed Tomography, Cardiac Nuclear Medicine Study, 2D Echo</td>
</tr>
<tr>
<td>History and Progress</td>
<td>Chief Complaint, Past Medical History, Allergy</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>Hypertension, Diabetes Mellitus, Atherosclerotic Disease, Obesity</td>
</tr>
<tr>
<td>Implants and Procedures</td>
<td>Valve Replacements, Coronary Angioplasty, Implantable Cardioverter-Defibrillator, Implantable Pacemaker, Aneurysm Surgery, Cardiac Resynchronization Therapy</td>
</tr>
</tbody>
</table>

Table 2. Common data elements for weight management/obesity (WM/O)

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms</td>
<td>Obesity, Overweight, Morbid Obesity, Abnormal Weight Gain, Obesity By Adipocyte Growth Pattern, Hyperplastic Obesity</td>
</tr>
<tr>
<td>Encounter Information</td>
<td>First Outpatient Appointment Date, Days Since Symptom Onset, Date Admitted, Date Discharged</td>
</tr>
<tr>
<td>Other Information</td>
<td>Height, Weight, Body Mass Index, Health Status, Disability, Smoking Status, Nutrition, Alcohol, Physical Activity</td>
</tr>
<tr>
<td>Laboratory Tests</td>
<td>Triglyceride, Glycerol, Cholesterol, Apolipoprotein, Lipoprotein, High-Density Lipoprotein, Low-Density Lipoprotein, Homocysteine, C-Reactive Protein, Thyroid Function, Liver Function Tests, H. Pylori Testing, Fasting Glucose, Hemoglobin A1c, Lipids, Serum Creatinine, Vitamin-D</td>
</tr>
<tr>
<td>Medications</td>
<td>Nonsteroidal Anti-Inflammatory Drug, Phentermine, Contrave Naltrexone, Contrave Bupropion, Qsymia, Belveq, Xenical, Metformin, Statins, Diethylpropion, Phendimetrazine, Benzphetamine, Liraglutide, Probiotics</td>
</tr>
<tr>
<td>History and Progress</td>
<td>Chief Complaint, Past Medical History, Allergy</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>Hypertensive Disorder, Diabetes Mellitus, Hyperlipidemia, Obstructive Sleep Apnea, Cardiovascular Disorder, Intracranial Hypertension, Depressive Disorder, Binge Eating, Degenerative Arthritis, Congestive Heart Failure, Nonalcoholic Steatohepatitis, Cancer, Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>Surgical Procedures</td>
<td>Bariatric Surgery, Laparoscopic Surgery, Gastric Bypass, Roux-En-Y, Lap-Band, Gastroplasty, Sleeve Gastrectomy, Duodenal Switch</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>Knee Arthroplasty, Cholecystectomy, Aortocoronary Bypass</td>
</tr>
<tr>
<td>Vital Signs</td>
<td>Temperature, Blood Pressure, Pulse, Respiratory Rate, Pain</td>
</tr>
<tr>
<td>Demographic Information</td>
<td>Address, Languages Spoken, Socioeconomic Status</td>
</tr>
<tr>
<td>Enrollment in Care-Coordination</td>
<td>Home Telehealth Monitoring</td>
</tr>
</tbody>
</table>

1882
Table 3. Common data elements for Kawasaki disease (KD)

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encounter Information</td>
<td>Fever Days at Admission, Date Admitted, Date Discharged</td>
</tr>
<tr>
<td>Other Information</td>
<td>Height, Weight, Age</td>
</tr>
<tr>
<td>Laboratory Tests</td>
<td>Erythrocyte Sedimentation Rate, C Reactive Protein, White Blood Cell, Hemoglobin, Platelet, Absolute Neutrophil Count, Gamma-Glutamyl Transeptidase, Alamine Transaminase, Albumin, Electrolytes, Urimalunysis</td>
</tr>
<tr>
<td>Imaging Tests</td>
<td>Echo, Cardiac Magnetic Resonance Imaging, Angio Computed Tomoraphy</td>
</tr>
<tr>
<td>Medications</td>
<td>Steroids, Intravenous Immunoglobulin, Naprosyn, Antiplatelets-Abciximab, Acetylsalicylic Acid, Clopidogrel, Anticoagulants-Heparin, Warfarin, Enoxaparin, Direct Thrombin Inhibitor, TNF-Alpha Antagonists-Infliximab</td>
</tr>
<tr>
<td>Echo</td>
<td>Left Main Coronary Artery, Left Anterior Descending, Right Coronary Artery, Left Circumflex Artery, Ejection Fraction</td>
</tr>
</tbody>
</table>

### Ensemble Methods

Our pipeline can integrate NLP tools using two basic ensemble methods: Union, which extracts the data element if any of the NLP tools detects the data element; and Intersection, which extracts the data element if both NLP tools detect the data element.

Although basic ensemble methods can combine the results of multiple NLP tools, the relationships (e.g., category, similarity, shared procedure or component) among data elements are not considered. To explicitly consider the relationships among data elements, we further converted the ensemble problem into a binary Multi-Label Classification (MLC) task, where each instance is a sentence in a note, features are the binary extraction results of each data element for each NLP tool, and labels are the binary ground truth of whether the data element exists in that sentence or not. For example, suppose there are 10,000 sentences in the data repository, 50 data elements to be extracted, and 2 NLP tools for the ensemble, then the corresponding MLC problem consists of 10,000 instances, 50 labels, and 100 (50 * 2) features.

To solve the MLC problem, we used five well-known algorithms to build our advanced ensemble: BR, which trains a binary classifier for each label independently; MLkNN, which is an instance-based method that extends the k-Nearest Neighbor (kNN) method to multi-label data; IBLR-ML, which is based on a formalization of instance-based classification as logistic regression and takes the correlation among labels into account; RAkEL, which randomly selects subsets of the label powerset (treating each distinct combination of labels as a different class) and combines them using a voting scheme; and ECC, which trains several classifiers in random order on a random subset of data, and combines them by voting. All advanced methods, except BR, learn relationships among labels (data elements) during training process.

### Test Corpus

We tested the above mentioned NLP tools and ensemble methods on 130 clinical notes (100 from the public domain and 30 private).

- **For public** clinical notes, TTK, PR and CNH collected 45,136 notes from MT Samples, i2b2 Challenges 2006, 2008 - 2012, 36, 37, 38, 40, 41, 42, 43 and ShARe CLEF eHealth Tasks 2013 Task 1 and 2, and 2014 Task 1, 44, 45, 46. Then, for each condition, the notes were selected based on keyword combinations: we used “congestive heart failure” for CHF, “weight management AND obesity” for WM/O, and “Kawasaki OR (fever AND rash AND red AND child) for KD. It should be noted that these rules only serve as simple filters for sampling notes, thus we did not consider synonyms or stemming. After narrowing down the search by keywords, we randomly selected notes for the evaluation from the filtered notes: 33 notes for CHF, 34 notes for WM/O, and 33 notes for KD. The total number of public domain notes was 100.

- **For private** clinical notes, TTK and CNH randomly sampled 30 notes, from a pool of 381 notes collected by JDC for KD (from Rady Children’s Hospital and the Emory University; Institutional Review Boards approved this study).
For the 130 notes, PR and CNH manually annotated 6,914 mentions of data elements (1,885 for CHF, 1,728 for WM/O, 1,678 for KD-Public, and 1,623 for KD-Private). TTK and CNH applied Stanford CoreNLP\textsuperscript{10} to split the 130 notes into 9,320 sentences (3,045 for CHF, 1,778 for WM/O, 2,824 for KD-Public, and 1,673 for KD-Private). We tested a total of four datasets: CHF (33 notes), WM/O (34 notes), KD-Public (33 notes), and KD-Private (30 notes). For each dataset, we randomly selected 50% of these notes for training, and held out the remaining 50% for performance reporting. We used 10-fold cross validation in training.

**Evaluation Metrics**

We considered a data element extraction to be correct using two different levels: corpus-level and sentence-level. For corpus-level, the prediction of a data element in a clinical note is considered correct if the data element does appear in the ground truth annotations in the same note (not necessarily in the same sentence). For sentence-level, the prediction is considered correct if the data element does appear in the ground truth annotations in the same sentence. It should be noted that the predictions are binary for both levels, thus multiple concepts of a corrected predicted data element would only count as a single true positive.

We computed Precision, Recall and F1-Scores for each method as our evaluation measures, and computed the corpus-level and sentence-level of each metric. The corpus-level F1-Score is defined as $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, where corpus-level precision $P = \frac{\# \text{of correctly predicted data elements}}{\# \text{of all predicted data elements}}$, and corpus-level recall $R = \frac{\# \text{of all correctly predicted data elements}}{\# \text{of all ground truth data elements}}$. The sentence-level F1-Score is defined as $\frac{\sum_{i} F_{i}}{N}$, where $N = \# \text{of sentences}, F_{i} = \frac{2 \times P_{i} \times R_{i}}{P_{i} + R_{i}}$ for each sentence $i$, sentence-level precision $P_{i} = \frac{\# \text{of correctly predicted data elements in sentence } i}{\# \text{of predicted data elements in sentence } i}$, and sentence-level recall $R_{i} = \frac{\# \text{of correctly predicted data elements in sentence } i}{\# \text{of ground truth data elements in sentence } i}$.

**Implementation**

We utilized the implementation of the MLC algorithms available in the MULAN\textsuperscript{11} package for ensembles, and applied J48\textsuperscript{12} as the base learner for the MLC algorithms. For cTAKES,\textsuperscript{3} we utilized the Dictionary Lookup Fast Pipeline\textsuperscript{5} and the built-in concept dictionary, which was a subset of UMLS\textsuperscript{13} containing SNOMED-CT,\textsuperscript{14} RxNorm,\textsuperscript{15} and all of the synonyms. The system was implemented in Java, Python, and Shell Scripts. Also, we released the ensemble component in our pipeline. In this component, there were three inputs: (1) ground truth annotations in BRAT format, (2) annotations generated by an individual NLP tool (also in BRAT format), and (3) the beginning and ending position of each sentence in notes (generated by the sentence splitter). This component can perform basic and advanced ensembles, compare the ground truth annotations, and output the corpus- and sentence-level evaluation results. The code is available at https://github.com/tsungtingkuo/ensemble.

![Figure 2](image-url)  
**Figure 2.** Corpus-level F1-Scores for NLP tools, basic ensemble, and advanced ensemble methods.
Results

Figures 2 and 3 show corpus- and sentence-level F1-Scores for the four datasets using the two NLP tools, two basic ensemble methods, and five advanced ensemble methods. The detailed Precision (P), Recall (R), and F1-Scores (F) for corpus- and sentence-level results are shown in Tables 4 and 5. In general, the scores of corpus-level evaluation are higher than those of sentence-level. We believe this is due to the aggregation of corpus-level results from sentence-level results, as sentence-level extraction is more challenging than corpus-level extraction.

For basic ensemble methods, Union generally improved performance over a single NLP tool, indicating that coverage is a critical concern for these data element extraction tasks (Figures 2 and 3). The importance of coverage can also be seen in Tables 4 and 5, where single NLP tools show high precision but relatively low recall. This also explains why Intersection consistently performs worse than all other methods (even worse than single NLP tools).

For advanced ensemble methods, although no method consistently performed the best, we did observe an interesting phenomenon: these MLC ensemble algorithms boosted performance on WM/O and KD-Private datasets, but they did not perform well on CHF and KD-Public datasets. This is related to label density, which is the number of ground truth annotations per sentence. For example, for WM/O the label density was (1,678 annotations) / (1,778 sentences) = 0.94, while for CHF the label density was only (1,885 annotations) / (3,045 sentences) = 0.62. Since the advanced ensemble methods applied multi-label learning, the datasets with higher label density (WM/O and KD-Private) provided better training examples for the classifiers and better recall (as shown in Tables 4 and 5). It should be noted that although we only use two NLP tools in our experiment, advanced ensemble methods may still be very useful to improve the extraction results. For example, consider we are extracting CHF data elements from this sentence: “Mr. X is being discharged on Lasix, Digoxin and Toprol daily.” One NLP tool may successfully extract “Lasix” as the data element “Diuretics,” while the other tool may successfully extract “Digoxin” as “Antiarrhythmics,” but both tools may fail to extract “Toprol” as “Beta-Blocker.” In this scenario, neither union nor intersection can improve the extraction results. However, if these three data elements ("Diuretics," "Antiarrhythmics," and "Beta-Blocker") are usually mentioned together in the training clinical notes, MLC ensemble algorithms may be able to recover “Beta-Blocker” even if both NLP tools miss the mention “Toprol.” We believe this is the reason why advanced ensemble methods are able to largely improve the results compared to individual NLP tools.

Also, the performance of the advanced ensemble methods is bounded by the limited availability of the annotated clinical notes. We anticipate that, when more annotated clinical notes become available for training and testing, the performance improvement of the advanced ensemble methods will be more consistent and obvious, especially for data with high label density. Additionally, in some settings (such as clinical notes from primary care providers with patients facing multitudes of conditions), the number of data elements might be on the order of hundreds of thousands, and thus more training examples are required for learning MLC models.
We also conducted qualitative analysis of our results. For each dataset, the data elements with highest and lowest F1-Score, extracted using the Union ensemble method (because it consistently performed better), are listed in Table 6 to illustrate which data elements were the most or the least successful in the extraction. A comparison of data elements with the highest and lowest F1-Scores suggests that, in general, items in the category history and progress are harder to accurately extract than those in comorbidities. However, for some categories such as medications and laboratory tests, the extraction performance varies for each data element. This observation also indicates that adding more diverse tools (e.g., specifically designed to extract history and progress, or to extract the data elements of medications or laboratory tests) may further boost overall performance. Our average processing time (seconds per note) is 1.14 for cTAKES, 44.29 for MetaMap, 0.19 for basic and 2.10 for advanced ensembles.

Table 4. Corpus-level Precision (P), Recall (R) and F1-Score (F) of NLP tools, basic ensemble, and advanced ensemble methods. Numbers highlighted as blue underlined text indicate the best scores for each evaluation trial.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Congestive Heart Failure (CHF)</th>
<th>Weight Management / Obesity (WM/O)</th>
<th>Kawasaki Disease - Public (KD-Public)</th>
<th>Kawasaki Disease - Private (KD-Private)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>NLP Tool</td>
<td>cTAKES</td>
<td>.882</td>
<td>.285</td>
<td>.431</td>
<td>.619</td>
</tr>
<tr>
<td></td>
<td>MetaMap</td>
<td>.842</td>
<td>.102</td>
<td>.181</td>
<td>.902</td>
</tr>
<tr>
<td>Basic Ensemble</td>
<td>Union</td>
<td>.890</td>
<td>.307</td>
<td>.456</td>
<td>.634</td>
</tr>
<tr>
<td></td>
<td>Intersection</td>
<td>.809</td>
<td>.080</td>
<td>.146</td>
<td>.916</td>
</tr>
<tr>
<td></td>
<td>Binary Relevance (BR)</td>
<td>.794</td>
<td>.317</td>
<td>.453</td>
<td>.738</td>
</tr>
<tr>
<td></td>
<td>Multi-Label K-Nearest Neighbor (MLkNN)</td>
<td>.797</td>
<td>.199</td>
<td>.318</td>
<td>.715</td>
</tr>
<tr>
<td>Advanced Ensemble</td>
<td>Instance-Based Logistic Regression for Multi-Label (IBLR-ML)</td>
<td>.370</td>
<td>.321</td>
<td>.344</td>
<td>.557</td>
</tr>
<tr>
<td></td>
<td>Random k-Labelsets (RAkEL)</td>
<td>.836</td>
<td>.313</td>
<td>.455</td>
<td>.737</td>
</tr>
<tr>
<td></td>
<td>Ensemble of Classifier Chains (ECC)</td>
<td>.431</td>
<td>.321</td>
<td>.368</td>
<td>.729</td>
</tr>
</tbody>
</table>

Table 5. Sentence-level Precision (P), Recall (R) and F1-Score (F) of NLP tools, basic ensemble, and advanced ensemble methods. Numbers highlighted as blue underlined text indicate the best scores for each evaluation trial.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Congestive Heart Failure (CHF)</th>
<th>Weight Management / Obesity (WM/O)</th>
<th>Kawasaki Disease - Public (KD-Public)</th>
<th>Kawasaki Disease - Private (KD-Private)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>NLP Tool</td>
<td>cTAKES</td>
<td>.354</td>
<td>.268</td>
<td>.293</td>
<td>.224</td>
</tr>
<tr>
<td></td>
<td>MetaMap</td>
<td>.132</td>
<td>.077</td>
<td>.091</td>
<td>.125</td>
</tr>
<tr>
<td>Basic Ensemble</td>
<td>Union</td>
<td>.369</td>
<td>.280</td>
<td>.306</td>
<td>.237</td>
</tr>
<tr>
<td></td>
<td>Intersection</td>
<td>.115</td>
<td>.065</td>
<td>.076</td>
<td>.101</td>
</tr>
<tr>
<td></td>
<td>Binary Relevance (BR)</td>
<td>.329</td>
<td>.281</td>
<td>.285</td>
<td>.791</td>
</tr>
<tr>
<td></td>
<td>Multi-Label K-Nearest Neighbor (MLkNN)</td>
<td>.268</td>
<td>.200</td>
<td>.211</td>
<td>.673</td>
</tr>
<tr>
<td>Advanced Ensemble</td>
<td>Instance-Based Logistic Regression for Multi-Label (IBLR-ML)</td>
<td>.211</td>
<td>.273</td>
<td>.209</td>
<td>.683</td>
</tr>
<tr>
<td></td>
<td>Random k-Labelsets (RAkEL)</td>
<td>.350</td>
<td>.276</td>
<td>.294</td>
<td>.790</td>
</tr>
<tr>
<td></td>
<td>Ensemble of Classifier Chains (ECC)</td>
<td>.325</td>
<td>.278</td>
<td>.282</td>
<td>.778</td>
</tr>
</tbody>
</table>
Table 6. Results for data elements with highest and lowest F1-Scores using the Union ensemble method. The data elements are ordered according to their F1-Scores (or the number of matches to break a tie).

<table>
<thead>
<tr>
<th>F1-Score</th>
<th>Dataset</th>
<th>Category</th>
<th>Data Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest</td>
<td>Congestive Heart Failure (CHF)</td>
<td>Medications</td>
<td>Heparin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comorbidities</td>
<td>Diabetes Mellitus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laboratory Tests</td>
<td>Serum Albumin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Surgical Procedures</td>
<td>Roux-en-Y</td>
</tr>
<tr>
<td>Lowest</td>
<td>Weight Management / Obesity (WM/O)</td>
<td>Medications</td>
<td>Acetylsalicylic Acid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comorbidities</td>
<td>Anticoagulants-Heparin</td>
</tr>
<tr>
<td></td>
<td>Kawasaki Disease – Public (KD-Public)</td>
<td>Laboratory Tests</td>
<td>Intravenous Immunoglobulin</td>
</tr>
<tr>
<td></td>
<td>Kawasaki Disease – Private (KD-Private)</td>
<td>Medications</td>
<td>Gamma-Glutamyl Transpeptidase</td>
</tr>
<tr>
<td></td>
<td>Kawasaki Disease – Public (KD-Public)</td>
<td>Laboratory Tests</td>
<td>Albumin</td>
</tr>
<tr>
<td></td>
<td>Kawasaki Disease – Private (KD-Private)</td>
<td>Medications</td>
<td>Intravenous Immunoglobulin</td>
</tr>
<tr>
<td></td>
<td>Congestive Heart Failure (CHF)</td>
<td>Laboratory Tests</td>
<td>Angiotensin-Converting Enzyme Inhibitor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>History and Progress</td>
<td>Warfarin</td>
</tr>
<tr>
<td></td>
<td>Weight Management / Obesity (WM/O)</td>
<td>Laboratory Tests</td>
<td>Warfarin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>History and Progress</td>
<td>Clopidogrel</td>
</tr>
<tr>
<td></td>
<td>Kawasaki Disease – Public (KD-Public)</td>
<td>Imaging Tests</td>
<td>Angio Computed Tomography</td>
</tr>
<tr>
<td></td>
<td>Kawasaki Disease – Private (KD-Private)</td>
<td>Imaging Tests</td>
<td>Echo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Terms</td>
<td>Kawasaki Disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laboratory Tests</td>
<td>Platelet</td>
</tr>
</tbody>
</table>

Conclusion

We developed an NLP ensemble pipeline to extract data elements from clinical notes, using state-of-the-art NLP tools and ensemble methods. We tested our pipeline on public and private notes for CHF, WM/O and KD. The results indicate that the Union ensemble method provides consistent improvement, while the MLC-based ensemble methods may be useful on datasets with higher density of concepts in the clinical notes. However, our study was limited to data elements selected for three cohorts only and limited by the size of clinical notes. Additionally, we used ensembles of only two NLP tools, thus our current evaluation results might not generalize to other pairs of systems nor to larger sets of systems. Nevertheless, we demonstrated that our proposed ensemble pipeline can improve the performance of NLP tools, and it may provide a practical solution for the extraction of data elements from clinical notes.

In the future, we plan to test our pipeline on a larger number of clinical notes, include additional data elements for these three and additional cohorts to increase generalizability, wrap additional NLP tools into the pipeline, and add more concept dictionaries to improve coverage (e.g., dictionaries that include acronyms). Furthermore, we will add additional ensemble methods to improve the overall extraction accuracy and sensitivity of our pipeline (adding more NLP tools enables us to use voting and include more complex ensemble algorithms). We also plan to perform additional analysis to explain the differences in the corpora and the performance of the ensemble methods (e.g., learning curve, execution time, and tool-category performance analysis), adopt parallel computing to increase scalability, and release annotations for the public domain clinical notes. We will disseminate our methods to other researchers, and test our processes in other clinical data research networks like pSCANNEr.

Acknowledgements

This work is funded by PCORI contract CDRN-1306-04819. We thank the following domain experts for identifying and refining the data elements for CHF, WM/O and KD: Howard Taras, MD, UCSD; Zhaoping Li, MD, UCLA; Jane Burns, MD, UCSD; Adriana Tremoulet, MD, UCSD; Michael Ong, MD, PhD, UCLA; and Paul A. Heidenreich, MD, MS, Stanford and Palo Alto VA. Part of the de-identified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY. The computational infrastructure was provided by the iDASH National Center for Biomedical Computing funded by U54HL108460 and managed by the Clinical Translational Research Institute CTSA Informatics team led by Antonios Kourtes, PhD, funded in part by UL1TR001442.
References

35. MT Samples, http://www.mtsamples.com
40. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. JAMIA. 2010.
50. Kim Y, Riloff E. A stacked ensemble for medical concept extraction from clinical notes. AMIA Jt Summits Transl Sci Proc. 2015.
Plexlines: Tracking Socio-communicative Behaviors Using Timeline Visualizations

John Lee, MS, Ha-Kyung Kong, Sanny Lin, Karrie Karahalios, PhD
University of Illinois Urbana Champaign, Champaign, IL, United States

Abstract

In this paper, we visualize children’s coordinated gaze, gesture, and vocalization to better understand communicative behaviors and to identify developmental delay, specifically in the domain of Autism Spectrum Disorders. To date, existing behavioral data from clinical assessment instruments are often stored in raw text files or spreadsheets. This wealth of data is then represented as a single number summarizing behavior. Our approach transforms this data into a graphical story of a child’s behavior. To do this, we created Plexlines, a graphical record of a child’s social and communicative behavior. When presented with Plexlines, clinicians and researchers formed their own strategies for exploring the visualizations and independently identified children in need of further evaluation. Feedback showed that Plexlines has the potential to be integrated into existing behavioral evaluation processes, aid in the detection of developmental delays in young children, and serve as a visual artifact to better communicate with parents.

Introduction

In cases of Autism Spectrum Disorder (ASD), early intervention—as young as six months old—is critical to successful treatment and can mitigate developmental delays [1]. However ASD is difficult to diagnose early, as there are no easy medical tests to diagnose the disorders, and each patient needs individual attention. Often times social and language delays do not show until children interact with their peers in preschool [2], which is considered a late detection and reduces the effectiveness of treatment [3]. Researchers and clinicians are now turning to quick screeners for early assessment that observe the communicative behavior in young children [4]. However, it is difficult for clinicians to make an accurate behavioral model based on data in its current text-based format. We developed Plexline as a graphical representation of behavioral data for use in understanding communicative behavior and spotting developmental delay in children under 30 months of age.

We seek to reveal the hidden layers of human behavior in a visual format, ready for rapid interpretation. Current clinical interfaces include charts and tables, but they do not provide a full picture for making behavioral assessments quickly. Instead of following complicated coding guidelines and instruments to produce a number that represents a level of social engagement, Plexline combines multiple data types to reveal the unique dynamics of a child’s behavior that cannot be observed first hand. For our visualizations, we gathered behavioral data from annotations of the Rapid Attention Back and Forth Communication Test (RABC) conducted by our collaborators. The RABC is a structured experimental 5-stage social play protocol (greet, ball, book, hat, and tickle) between a child and an examiner. It was developed jointly by Emory University and the Georgia Institute of Technology for children aged 9-30 months to collect data about a child’s social and communicative behavior as a pre-screener for ASD [5].

Using hand-coded annotations, we categorize RABC communicative behaviors into gaze, gesture, and vocalization. The categories are color-coded and each behavioral event is drawn on the timeline as a colored circle. Laying out these circles along a timeline creates an overview of an individual child’s behavior. By taking advantage of a common visual language, we reveal complex patterns of coordinated behavior.

Two Plexlines

In this section, we introduce a use case for Plexlines by presenting two contrasting Plexlines from two different children (see Figure 1). A legend for Plexlines can be seen in Figure 1c. In this stage of the RABC, the examiner presents a ball to the child and says, “Look at my ball.” Typical reactions to the examiner’s statement (also referred to as an examiner bid) include looking at the examiner (blue circle) as they are speaking, and using either gesture (green circle) or speech (red circle) to ask for the ball. Figure 1a shows the Plexline of a child that display this common pattern of behavior. The diameter of the circle is proportional to the duration of the annotated behavior.

There is a variance in the timing and behavior responses in the Plexlines of typical children. However, the Plexlines share a common rhythmic pattern of overlapping circles representing coordinated gaze, gesture, and vocal behaviors. Notice how the Plexline of the second child in Figure 1b compares to that of the first child. Unlike the first Plexline,
the green circles for the gestures are small and there is lack of blue circles representing gaze towards the examiner. When watching the corresponding video for this Plexline, the child appears sociable to an untrained eye. However, the child fixates on the ball and rarely makes eye contact with the examiner. This lack of interest in the examiner is a critical sign of developmental delay [6, 7].

Plexlines allow users to quickly spot outliers in communicative behaviors such as those exhibited by the second child, and capture nuances in behavior that are difficult to spot in person or in video. We will provide more detail on the design of the Plexlines and the tool for manipulating Plexlines later in this paper.

![Figure 1. Two Plexlines showing different levels of communicative engagement during the ball stage of the RABC. In (a), the child’s gestures (green) are punctuated by gazing at the examiner’s face (blue). In (b), the child shows no gaze toward the examiner, a warning sign for developmental delay. (c) is the legend of the annotation categorizations. An Examiner Bid is an attempt by the examiner to initiate a response. The blue gaze circle denotes gaze towards the examiner. The green gesture circle denotes any communicative gesture. The red speech circle denotes vocalization and verbalization. Hatch marks represent other non-directed behaviors.](image)

Related Work

Current work in behavioral visualization is often limited to one aspect of behavior, such as visualization of only speech or only eye gaze (e.g. [8-10]). In addition, there is little application of behavioral visualization in current clinical processes. We tackle the challenge of visualizing multimodal behavior on a temporal graph. In this section, we will discuss our main inspirations from visualization work in healthcare, though this is not an exhaustive list of work in the field.

Prior work in visualization in healthcare focuses on electronic health records. Rind et al. have reported on the extensive literature in information visualization for electronic health records in their recent work [11]. The ongoing work of LifeLines2 [12] and EventFlow [13] have produced powerful tools for visualizing sequences of events, notably for visualizing a patient’s treatment history. Hsu et al. built a tool which generates a timeline interface showing the change of medical condition of patients to understand and summarize patients’ conditions [14]. Ozturk et al. showed that visualizing a patient’s medication history can be helpful, especially in emergency care, and developed an application that converts a patient prescription history data into a simple timeline visualization [15]. Visualization of medical records not only aid doctors in presenting data in an efficient manner, but also in discovering insights. Klimov et al. and Shahar et al. both focused on building an intelligent interactive visualization system which dynamically presents relevant clinical data to the clinicians by utilizing existing domain knowledge [16, 17]. Work by Gotz et al. visualized patient data based on similarity to allow human experts to refine and reference comparable patient data [18]. Similarly, Stubbs et al. proposed an interactive system for exploring and visualizing data to identify similar patients from a database [19].

While there is extensive work in visualizing personal medical histories, visualizing behavioral data is underrepresented in this field, especially for diagnosing and pre-screening for ASD. However, more researchers are applying similar technology for screening and diagnosing ASD to supplement existing procedures. Boraston et al. have utilized eye-tracking behavior to investigate gaze behavior [20], and Hashemi et al. used computer vision to assess visual attention in children [21]. More recently, Han et al. presented a visualization that allows for a comparison of any two behaviors in the RABC dataset that occur at or around the same time across multiple sessions [22]. In our work, we focus on presenting many co-occurring events of a single child in a compact visualization, which allows clinicians and researchers to examine and compare behavioral records of multiple children.
Project Background and Motivation

We focus on techniques for behavioral imaging of social and communicative behaviors and creating graphical visualizations to aid clinicians in making a diagnosis. Currently, researchers and clinicians often collect quantified scores of various behaviors from the behavioral coding stage. Well-known protocols such as the Autism Diagnostic Observation Schedule (ADOS), or the Autism Diagnostic Interview-Revised (ADI-R) result in a single number or binary evaluation representing social communicative engagement with no convenient way to interpret and explore behavior data temporally [23, 24].

In order to create graphical visualizations of child behavior, we have collected data of over 100 children, aged 9 to 30 months, engaged in a five-minute RABC session. The examiner evaluates the child’s responses to the explicit social bids and the ease of engaging the child. Specifically, the examiner seeks to elicit social attention, back-and-forth interaction, and social communication from the child. The dyadic nature of the RABC allows us to break interactions down into smaller parts and observe engagement at different levels of granularity. Our collaborators collected the RABC data as part of a NSF Expeditions effort and annotated each RABC video. Via a training codebook, three independent coders were trained to code the RABC sessions. After reaching at least 90 percent overlap for each of the annotations, they proceeded to annotate the remaining video corpus frame by frame, categorizing the child’s gaze, gesture, and speech.

Annotation Taxonomy

As a human-centered timeline, Plexlines are styled to focus on the child, with brightly colored circles representing the child’s socio-communicative behaviors. We are mainly interested in the child’s reactions and responses to the examiner bids. We separate the social behaviors of the child into gaze, gesture, and speech, which are color-coded blue, green, and red, respectively. Because we are primarily interested in behavior directed toward the examiner, only such behaviors are included in the default circle categories. Thus, an annotation labeled gaze at the examiner is visualized as a blue circle, while gaze at the ball is not. However, we still include the annotations of the non-directed behaviors in the visualization for reference. Non-directed behaviors are grouped into an independent category and visualized as hatch marks on the Plexline, regardless of type of behavior (see Figure 2c).

The examiner’s speech is categorized as social bids or verbalization. They are usually social bids, which consist of a question or a demand, such as “Can you turn the page?” or “Look at my hat!” Social bids have a limited range of expected reactions and are marked by small black points on the Plexline (see Figure 2b). They pop out among the colored circles and indicate the start of an interaction. All other verbalizations (except when calling the child’s name) by the examiner are visualized as small, unfilled black circles (see Figure 2c).

Visual Design

Plexlines’ abstracted approach of using shapes and symbols can be more concrete than numbers. People perceive perceptual structure—patterns of quantities, color, relative size, and shapes—better than patterns of numbers or words [25]. The latter uses single units, while the former operates as a whole through our cognitive ability to quickly perceive spatial relationships.

Through an iterative design process [26], the design of Plexlines was adapted to be heavily informed by the patterns clinicians look for in behavioral coding. Based off of the annotation taxonomy, Plexlines use overlapping circles along a timeline to highlight coordinated behaviors. The temporal nature of the sessions makes it natural to lay the information out horizontally as a timeline [27]. This layout also facilitates comparing a list of Plexlines.

Coordinating Co-occurring Behaviors

The coordination of these three behaviors indicates high social engagement. It is one of the motives behind the choice to use circles to represent the child’s behavior. To show coordinated behavior on a timeline, the design must...
accommodate overlapping behaviors. After iterating through many visual styles, slightly transparent circular shapes were found to be the most legible for identifying overlapping behaviors with minimal occlusion.

The diameter of the circle scales relative to the duration of the child’s behavior. Although this creates a non-linear scale between the area of the circle and the duration of the annotation, exponential scaling assists users in identifying co-occurring behaviors quickly to then explore in more detail. In cases where circles overlap, the smallest circle is always brought to the top to minimize occlusion, regardless of color. Circles are also transparent to allow for ease in spotting overlaps. A clinician interested in complex coordinated behaviors would look for instances of two, or even three, circles of different colors overlapping or lack thereof (see Figure 2c). Examiner speech throughout the RABC is consistent and short, so the circles for the examiner do not scale to duration. They act as anchor points along the Plexline for comparison across multiple Plexlines.

Defining Initiation-Response Stages

Joint attention - characterized by the shared focus of two or more people on one object - is another critical measure of social engagement [28]. The fluidity of back-and-forth interactions is one way to measure joint attention. Much of the RABC centers engagement around objects—a ball, book, and hat. When the examiner presents the object to the child, she is probing for signs that the child is able to identify her intent. Ideally, the child will draw attention to the object by means of gaze, pointing, or verbal behavior to create a shared experience between the child and the examiner.

Examiner bids followed by a series of colored circles (or lack thereof) before reaching the next social probe may be isolated as one initiation-response sequence. One initiation-response sequence shows how soon and how much the child responds to the examiner. A child-examiner pairing that is highly synchronous would have a rhythmic pattern of initiation-response sequences along the entire Plexline (see Figure 1a). The Plexline of a distracted or object-centered child would have extended breaks throughout the Plexline (see Figure 1b).

Building and Exploring A Plexlines Library

In demonstrating Plexlines to our audience, we do it in a browser-based environment that allows them to compare multiple Plexlines (see Figure 3). Facilitating comparison is critical in establishing the baseline metrics and demographic profiles necessary to spot deviations. The interface follows Shneiderman’s mantra of “overview first, zoom and filter, then details-on-demand” [29]. Users start by viewing a single child against an archive of all the Plexlines laid out as small multiples for simple comparison [30]. While the visualization is designed to stand alone without video, a video can be loaded with each Plexline. The Plexline acts as a seekbar for navigating the video which is provided to assist in learning and to clarify points of interest that may be confusing.

Figure 3. The main view of the webtool with one Plexline selected.
Comparing Apples to Apples

Behavioral development is highly dependent on age, especially in young children. For instance, we do not expect to see much vocalization in children less than 12 months old. Filtering the archive by age and gender in the webtool sets the expectations for what patterns of behavior are common among the age groups. We provide an aggregate view, where all other Plexlines—besides that of the child in focus—are layered on top of each other (see Figure 4). Patterns become even easier to compare when the user places one child’s Plexline against an aggregation of all of the other Plexlines in the filtered population of children.

As an exploratory tool, filtering and sorting by certain age groups confirms what we know about developmental milestones and helps identify different trajectories of behavioral development. Aggregation allows clinicians and researchers to create experimental “templates” of behavioral styles and develop predictive patterns of behavior for different populations of children. This is exemplified in the comment, “I really liked being able to look at one child’s performance compared to the [aggregate view] of all the others” (P5). All of our study participants commented on the value of the aggregate display (see Figure 4), with five commenting it as one of their favorite features. Clinicians further suggested we include aggregate templates for age ranges at three-month intervals and templates for specific behavioral disorders.

Figure 4. One Plexline compared against an aggregate of the same sequence of events.

Customization

Plexline uses predefined groups for 21 different annotations of child behavior. The directed behaviors are classified into color-coded circles, and the non-directed behaviors are displayed as hatch marks. While we initialize each category for users, the users can customize the Plexline by adjusting these categories. We provide users with a simple drag-and-drop interface that allows them to move the annotated child actions into other categories (see Figure 5). With this functionality, users can tailor Plexlines to their specific needs. If one would like to focus only on one specific child behavior such as pointing, all other behaviors can be removed from the Plexline to show a simplified version highlighting that one behavior. To visualize “gaze at ball” annotations as circles instead of hatch marks on a Plexline, a user can move “gaze at ball” from the others category to the gaze category through the annotation module. The flexibility of the webtool accommodates researchers and clinicians with distinct needs and provides an opportunity for users to delve deeper by showing less.

Figure 5. Customization and authoring in the webtool tool allows the user to change annotation categorizations and simplify the Plexline. This image shows three stages of filtering. From top to bottom: (a) all annotations visible, (b) only gestures visible, (c) only pointing gestures visible
Evaluation

To evaluate Plexlines, we recruited eleven (9 female, 2 male) researchers and clinicians with research and/or clinical experience in developmental health, including early childhood education, intervention, and autism research. Six of the participants reported that they had more than five years of research experience. The remaining participants were doctoral candidates.

We began the study with a five-minute introductory video, which explained the details of the RABC, Plexlines, and the webtool. The participants were then given approximately thirty minutes to familiarize themselves with the webtool. During this time, the participants explored the functionality of the webtool and described how they might use it in their workflow. We observed the participants using the tool and recorded notes describing their use of the webtool and Plexlines and any comments they shared while using it. After the subjects explored the webtool, they were given a Plexline comprehension worksheet, which asked the participants to describe the behaviors represented by four different Plexlines without the assistance of video or the webtool. Lastly, the participants were given a survey that included questions to evaluate Plexlines and the webtool, and indicate their knowledge of developmental health and ASD behaviors on a 5-point Likert scale. In addition, we included open-ended questions to list the strength and weaknesses of Plexlines and the webtool. The study lasted roughly one hour for all participants.

Despite our initial concerns that researchers and clinicians would be hesitant to embrace unfamiliar behavioral visualization technologies, Plexlines and the webtool were well received. On a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5), participants rated that they were satisfied with Plexlines ($\mu=4.11$, $\sigma^2=0.33$) and the webtool ($\mu=4.33$, $\sigma^2=0.24$). We found that the participants grasped the concept of Plexlines quickly, and they were able to make interpretations from the Plexlines.

We used an open-coding method to label and summarize comments and observations while participants were exploring the webtool. We discovered that in general, participants used the following strategies to navigate Plexlines and draw interpretations, going through each stage at different rates depending on their level of familiarity with ASD and their research goals:

1) Rapidly explore the archive through a browsable interface.
2) Compare one child against many using customizable zooming, filtering, and aggregation tools.
3) Interpret Plexlines by spotting specific moments of concern.
4) Use Plexlines as a narrative to share their interpretations of a child.

Exploring the Archive

Facilitated by the webtool interface, the participants initially browsed the library of Plexlines displayed as a list to get an overview of the tool and familiarized themselves with Plexlines. After spending a few minutes browsing, each of the participants in our study independently developed their own search plans for exploring the Plexlines. The goals varied depending on their previous research, clinical experiences, and familiarity with behavioral analysis. Some participants focused primarily on finding patterns among different age groups, while others focused on spotting specific behaviors in children, catered to their own research goals or interests.

Interpreting Plexlines

After the exploration stage, the webtool allowed users to customize the Plexlines display to focus on specific points of interest. One participant described her hypothesis and investigation strategy as such: “If you are working with children with ASD, the first thing you want to look [for] is gesture and gaze because they usually avoid eye contact and have fewer gestures compared to children with disabilities” (P10). Participants used relationship comparison and details-on-demand strategies to verify their hypotheses and to make interpretations about the child’s activity on the Plexline.

On a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5), participants found Plexlines interpretable ($\mu=4.22$, $\sigma^2=0.39$) and exhibited visual literacy of the Plexlines within the first thirty minutes. One participant stated, “I’ve been here 15 minutes and already [they’re] a lot easier to read” (P1). Every participant commented on atypical behaviors they observed throughout the Plexlines, and were able to independently pick out a child that was suspected to be at risk for autism. When spotting an outlier, the participants would make remarks that indicated they wanted to pursue a deeper investigation: “I can see that this guy is not responding to these bids much at all” (P4), “I really want to know what is happening with the little girl!” (P4), and “RA052 stands out the most”.
Most subjects then spent considerable time on one child, closely inspecting the corresponding Plexline and using it to scroll to the video to verify their hypotheses. The ability of the participants to identify children in need of further evaluation based solely on a graphic demonstrates the salience of Plexlines.

Participants also distinguished between the subtleties in interaction styles. There was consistent agreement among the participants about what types of behaviors the children were showing at specific times. Four participants also felt comfortable expressing concern for specific children based solely on the Plexline without any video or aggregation for comparison.

Making Comparisons

One popular strategy that the participants displayed was filtering by age and then searching for patterns within the age group. One participant described their process as follows: “Watching her, I immediately think something is going on. She is making a lot of gestures, she is making sounds, but it is not clear to me what is happening yet, so I am filtering down to a tighter age range” (P4). The ability to filter profiles and create aggregates for that group proved to be critical for providing the right context for making comparisons. Common filtering strategies include creating profiles by age or focusing on specific behaviors.

Participants were able to quickly glance over the small multiples to make comparisons against other children in the library. One participant wrote that one of the most positive aspects of Plexlines is “the ability to examine patterns across multiple subjects in one visual field” (P11). Beyond simple browsing, the aggregate view (see Figure 4) is a popular and effective method for making comparisons. Five participants specifically stated that the ability to view the aggregate was one of the biggest strengths of Plexlines in providing anchors for behavioral expectations. Being able to conduct a data analysis on the whole data set is especially critical in explaining visualizations to other people.

For researchers that were interested in a specific type of behavior, they preferred to filter by hiding all other behaviors besides the ones they are most interested in. The customizability of the annotation schedule allowed the participants to tailor the Plexlines to their expertise. One particular participant adjusted the settings to show only 3 gestures: point, reach, and tap. When asked to explain his process, he said, “I am interested in gesture because he is trying to say something, but he is not able to vocalize it” (P11). After spotting an unusual Plexline, participants formed hypotheses about what they thought was happening and confirmed their interpretations by focusing on specific stages of the Plexline.

Using Plexlines to Tell a Narrative

Visualizations can be made even more accessible by allowing the users to filter out noise and frame the information around a narrative [31]. The ability to omit data and break down behavior by type or stage proved to be valuable in telling a story to non-experts. After completing an evaluation session and interpreting the results, clinicians in our study wanted to engage a parent in the next stage decision-making process. In describing something as complicated as behavior, showing is more informative than telling.

As a colorful graphic, Plexlines are less intimidating than a spreadsheet or a number rating. Because of this quality, Plexlines act as a catalyst for discussions between clinicians and parents. A clinician can navigate a Plexline in part or in whole to tell a story with a beginning, middle, and end to the parent through the visualization. In an example situation that a participant gave us, clinicians would “use it to say, ‘Four different times I asked your child to do this and he didn’t respond. Our expectation is that with the first or second bid, [the child would respond]’” (P2). The ability of a graphic to show a causation of events and allow the viewer to discern the relationship among them creates a narrative that engages the reader beyond a text-only report.

To further reduce complexity for the parent, the clinician can break the session down into tasks by segmenting the Plexlines and filtering out unnecessary annotations. For example, parents may have misconceptions about their child’s developmental health. One participant gave an example of a parent that believed a highly vocal child indicated no developmental delay: “A dad linked [social engagement] to language. So let’s remove all the vocalization. Let’s [focus on] where the child is looking to show a parent” (P2). By isolating the visual narrative around eye gaze and having a concrete representation of behavior as an artifact for discussion, the clinician can more easily explain her interpretations.

Study participants described telling stories to clinicians, to parents and to research audiences. For example, a researcher stated, “Once I had done data analysis on the whole data set, I still might present [the] data aggregate and see [the] prototypical [18 months old child that was engaged] like for a talk. Then show a prototypical for a
child who is impaired” (P1). In this same context, the archival capabilities of the webtool allow users to create a comprehensive story that extends beyond just one child. According to a participant, Plexlines “display a visual story to others without worrying about confidentiality. Plexlines replaces the need of sharing videos” (P9).

Discussion and Future Work

Plexlines do not replace the existing evaluation processes, but can assist in understanding child behavior. In general, visualization is an underutilized technique in behavioral science, and Plexlines demonstrates the potential of this approach. From our user feedback, we found that Plexlines complement traditional behavioral evaluation processes, especially by providing salient visualization of social behavior. We identified opportunities to use Plexlines in rapid screening, tracking child development, and training.

Screening a Spectrum of Children

We intentionally did not create profiles of prototypical children because we realize that no two children are alike and there is no such thing as an “average” child. Visually comparing a child to their age group, or even just comparing a child to another is not a trivial task with existing tools. Currently, we suggest how a child might behave by showing the aggregate view, but this can be visually cluttered and misleading. The additive nature of transparent colors exaggerated the characteristics of a typical child, yet we discovered that comparing aggregates and averages was one of the most important exploration strategies, helping the users form frames of reference (see Figure 4). After a few minutes exploring the tool, participants familiar with RABC and had experience with ASD were comfortable stating, “Here’s a good example of a typical kid” (P1).

While we were initially hesitant to create any Plexlines that are suggestive of an ideal child for a specific profile, the participants’ desire for a normative graph with normative data changed our perceptions. A single normative Plexline that is informed by the dataset can be valuable in all stages of the process from exploration to interpretation to sharing. This is an opportunity for us to explore in the future.

Plexlines as a Personal Archive

While our user study focused on the Plexlines, there is room for improvement in making the interface more robust and usable. In particular, we plan on allowing the user to focus on and track the records of a single child. In our dataset, there are several children who were brought back for a follow-up evaluation. These children have two Plexlines in the library. These follow-up RABC sessions open up opportunities to track a child’s progress consistently over a longer period of time. This personal record can be used by parents and clinicians to build a cohesive, sharable snapshot of each child as they navigate the child’s future, from pre-screening for autism to evaluating the effectiveness of intervention strategies. Users will need the ability to mark and annotate points of interest on the Plexline for later reference. Each added Plexline contributes to a growing archive of Plexlines that help researchers and clinicians understand communicative child development.

In creating such a record, we are exploring Plexlines in other types of behavioral evaluation beyond the RABC. A few of our participants have shown interest in visualizing their own datasets featuring dyadic interaction with Plexlines. We are exploring alternate visualizations for dyadic interaction where one person is visualized above the center line and the other person is visualized below.

Plexlines as a Teaching Aid

Not only are Plexlines valuable for explaining the evaluation process and outcomes to parents, but they can also be used for training clinicians in the RABC. Trainees can see a series of Plexline visualizations to understand a typical session and use the exploration strategies to understand the structure of the protocol and the expected behaviors that follow. Used as an artifact in conversation, visualizations can help improve memory and support interpretations [30]. Several participants commented on the use of Plexlines as a teaching tool. On in particular emphasized that Plexlines can reveal patterns that are difficult to spot in videos: “I think this would be helpful [for training]. For example, Amy Whetherby’s videos for training say what’s typical and not typical. [Plexlines let you] see the whole thing - like a condensed ADOS. [I’d go] through an example of what I’m looking for: verbal, nonverbal, social non-directedness. It’s a hard skill to train” (P4).

Limitations

Our approach to visualizing behavior on Plexlines relies heavily on annotation availability. Our current RABC dataset may not contain many cases of children on the autism spectrum. Most children participate prior to a
diagnosis. Additionally, the annotations are limited by the accuracy of coding, and the rigidity of the handcrafted annotation schedule. We realize annotations are imperfect and not inclusive of every possible behavior.

Coding schemes improve iteratively over time. The features we are coding may not be optimized for our screening goals, and the taxonomy for Plexlines is highly influenced by the taxonomy of the existing RABC annotations. We currently distinguish between gaze, gesture, and speech behaviors. Smiles and affect are examples of annotations that may be useful, but not yet annotated. Annotations for other red flags of autism, such as echolalia, unusual prosody, and stereotypical repetitive behaviors [31, 7] could also be explored.

The time to code the RABC videos is another limitation. Ideally, a clinician may want to show a parent their child’s Plexline immediately following the session. Other members of the research team are exploring vision and audio techniques to automatically extract annotations from the RABC videos. Classification accuracy rates for gaze, speech-like vocalization, smiles, and emotion are improving. Until we can reliably annotate data automatically, we begin by using time-intensive hand-coded annotations for our visualization.

Conclusion

We presented Plexlines, a technique for visualizing multiple layers of socio-communicative behaviors on one timeline. The visual presentation of behavioral data provides researchers and clinicians with a novel way of understanding and interpreting behavioral communicative data beyond traditional charts and tables. While initially fearful that alternative visualizations were not the norm in clinical and research practice, feedback from researchers and clinicians shows that Plexlines serves as an engaging and effective method of providing a compact overview of a person’s behavior.

Our study revealed that Plexlines has the potential to be integrated into existing behavioral evaluation processes, aid in the detection of developmental delays in young children, and serve as a visual artifact to better communicate with parents. Through exploration and customization of Plexlines, researchers and clinicians were able to independently identify two children at risk for ASD. Participants used Plexlines to interpret the RABC data and reflect on their own personal research questions. They created stories around Plexlines and imagined using Plexlines to facilitate conversations with parents and larger audiences. Personalizable and shareable, Plexlines aid in expanding our shared understanding of human behavior and child development. While more studies are needed, we envision integrating Plexlines into early screening and clinical evaluations in hopes of increasing rates of early detection for timely intervention, a critical element in minimizing developmental delays among children with ASD [32].

References


Improving Endpoint Detection to Support Automated Systematic Reviews

Ana Lucic, MS¹, Catherine L. Blake, PhD¹
¹School of Information Sciences, University of Illinois, Champaign, IL

Abstract

Authors of biomedical articles use comparison sentences to communicate the findings of a study, and to compare the results of the current study with earlier studies. The Claim Framework defines a comparison claim as a sentence that includes at least two entities that are being compared, and an endpoint that captures the way in which the entities are compared. Although automated methods have been developed to identify comparison sentences from the text, identifying the role that a specific noun plays (i.e., entity or endpoint) is much more difficult. Automated methods have been successful at identifying the second entity, but classification models were unable to clearly differentiate between the first entity and the endpoint. We show empirically that establishing if head noun is an amount or measure provides a statistically significant improvement that increases the endpoint precision from 0.42 to 0.56 on longer and from 0.51 to 0.58 on shorter sentences and recall from 0.64 to 0.71 on longer and from 0.69 to 0.74 on shorter sentences. The differences were not statistically significant for the second compared entity.

Introduction

Scientific literature in the field of biomedicine continues to grow at a staggering rate. The number of abstracts in PubMed already exceeds 24 million and every week, the National Library of Medicine adds thousands of new abstracts. Although systematic reviews can help both frontline health care professionals and researchers by accurately synthesizing high quality evidence, the manual processes used to conduct a systematic review are time consuming. A systematic review, which is the cornerstone of Evidence Based Medicine (EBM), can take 5-6 people more than 1000 hours to complete¹, so help is urgently needed to reduce the time between the publication of new results and their integration into practice. Figure 1 shows clearly that the number of meta-analyses (a systematic review that integrates results using quantitative methods) on diabetes have increased from 40 to 974 since 2000.

Automating the systematic review process was first introduced almost a decade ago² and since then there have been several efforts to that focus on the information retrieval³-⁸, and information extraction stages of the process⁹ have been developed. In addition to automated strategies, manual efforts are underway to capture data required in a systematic review and tools are available to help with writing the manuscript. Our goal is to support the systematic review process by automatically identifying results from full-text articles.

In this paper we focus on comparison sentences. Within the context of biomedical collection of articles, comparison sentences frequently communicate the results of a study and include the information about the entities that were compared and the basis on which they were compared, which we call an endpoint. Although not a frequent structure in scholarly articles, comparison sentences contain a wealth of information that, when viewed in aggregate, can assist policy makers, health care providers, patients and general consumers of health information with insights about the entities of interest and their comparative characteristics. An analysis of clinical questions in the National Library of Health (NLH) Question Answering Service (http://www.clinicalanswers.nhs.uk) revealed that 16% of the 4,580 questions referred to direct comparisons of different drugs, treatment methods and intervention¹¹. Although comparisons have been identified as an information need, current systems do not allow the extraction and synthesis of comparative data from scholarly articles. More broadly, the structure of comparison sentences and the methods that allow parsing of the comparison structure in an automated way can be seen as particularly relevant to Comparative Effectiveness Research whose goal is to provide evidence on the effectiveness, benefits, and drawbacks of different treatment options. Identifying comparison facets in an automated way can assist the process of generating a comparative summary and thus highlight the areas where comparative work has or has not been done. More recently, there has been a shift towards the identification of indirect comparisons in scholarly articles¹².
High quality evidence consisting of systematic review of randomized clinical trials that provide direct (head-to-head) comparison of two interventions are commonly rare, sometimes non-existent or inconclusive; occasionally, indirect comparisons can be more reliable than direct evidence due to methodological inadequacies of trials.

In this study we pay particular attention to direct comparisons (direct mention of at least two compared entities and the endpoint in the comparison sentence) and to the expression of endpoints in comparison sentences. In the following example, fast-track and slow-track patients represent the compared entities whereas HbA1c, blood pressure and serum creatinine levels represent the basis on which slow-track and fast-track patients were compared. This sentence is considered a direct comparison sentence:

(1) HbA1c [Endpoint_1], blood pressure [Endpoint_2], and serum creatinine levels [Endpoint_3] were significantly higher in fast-track [Entity_1] than in slow-track patients [Entity_2].

The following is an example of a comparison sentence that features hypoglycemia as the endpoint modifier:

(2) Glucagon levels [Endpoint_1] were significantly lower (P < 0.0001) during hypoglycemia [Endpoint_1_modifier] with tolbutamide [Entity_1] than without tolbutamide [Entity_2].

In the above example, hypoglycemia is used to modify the main endpoint, glucagon levels, and to compare it with relation to drug tolbutamide.

Sentence (3) features body weight, as an endpoint:

(3) Body weight [Endpoint] of the high-fat-fed C57BL/6J mice [Entity_1] was 28% higher at 3 months (P < 0.001) and 69% higher at 15 months (P < 0.001) compared with the normal diet-fed C57BL/6J mice [Entity_2].

Sentence (4) features fasting plasma glucose (FPG) as an endpoint that indicates the difference that was observed in ZDF rats at 12 weeks of age versus 6 weeks of age:

(4) The fasting plasma glucose level [Endpoint] of ZDF rats [Entity_1] was significantly elevated at 12 weeks of age [Entity_1 modifier] compared with the level observed at 6 weeks of age [Entity_2 modifier].

These examples demonstrate that comparison sentences represent a convenient medium for examining and identifying endpoints reported in biomedical literature. Furthermore, this paper will demonstrate how by focusing on comparison sentences in biomedical articles we can facilitate a better identification, retrieval, aggregation as well as examination of endpoints. Once entities and endpoints are identified information from comparison sentences can be organized into a tabular summary that shows a detailed summary of which comparisons have already been made and which comparisons are currently missing from the current biomedical literature (see ref 18 for details). Such a summary can be used when writing a systematic review to establish areas where there is enough literature and to identify areas where the results between studies differ.

Several automated methods have been developed to identify comparison sentences from text and we extend that work by identifying the specific noun phrase within a sentence that fulfills each of the entity and endpoint roles. Previous work achieved good results with respect to Entity 2 where 0.74, 0.80, and 0.91 were reported for precision, recall and accuracy; however, differentiating between Entity 1 and the endpoint is challenging because they are used in similar contexts and with similar grammatical structures. This paper introduces a new method to improve the predictive accuracy of Entity 1 and the endpoint. The approach employs a set of heuristics that capture measurements, and leverages a multi-class classifier instead of a binary classifier. The hypothesis of this study is that endpoints frequently, although not exclusively, represent dependent entities that lend themselves to measurement. Also, we hypothesize that this property can be useful for separating endpoints from other entities in the sentence and for improving the precision and recall for this comparison facet.

Materials and Methods

A set of 100 comparison sentences from the journals Diabetes, Carcinogenesis, and Endocrinology (TREC Genomics collection) that comprise 641 noun phrases used in an earlier study were enriched with information about whether the head noun of the candidate noun is likely to be categorized as an Amount or whether it is more likely to be as a (population) group. This set of 100 comparison sentences was used for training the models (training set). A locally created dictionary consisting of 91 terms was used for the purpose of enriching the feature set with the information on whether the head noun of the candidate noun phrase in a comparison sentence was likely to be an...
Amount or a population group. 71 unique terms such as level, concentration, rate, mass, proportion, and degree were categorized as an Amount and 20 terms were used to indicate a population group such as, control, arm, trial, treatment. Drugs were also identified as a group because drugs, within the context of comparison sentences extracted from biomedical scholarly articles, are frequently used as a population group that is compared to another group of drugs. Drugs were identified using the UMLS Pharmacologic Substance semantic class. A number of candidate nouns that occur in comparison sentence will not be identified with either Amount or Group semantic class and was assigned a Null value.

Once the models were built (Table 1), they were then applied to the entire collection and evaluated on a test set of 132 sentences with <=40 words that were drawn at random from the collection. The test set comprised 939 noun phrases. Results are also shown for a sub-sample of the test set consisting of 66 short sentences (<=30 words) comprising 385 noun phrases18 to explore the impact of sentence length on system performance. Our research question is to determine if the additional amount and group information improves classification performance for the three crucial facets of a comparison sentence – the two compared entities and the endpoint on which they are compared (Entity 1, Entity 2, and Endpoint).

**Table 1.** Description of the six experiments. Support Vector Machine classification algorithm and linear and Gaussian kernel were used for all experiments.

<table>
<thead>
<tr>
<th>Experiment title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC (Binary classifier)</td>
<td>Support vector machine binary classification method, one versus all (OVA).</td>
</tr>
<tr>
<td>MC₃ (Multi-class classifier, 4 classes)</td>
<td>Support vector machine multi-class classification with no additional features (4 classes, Entity 1, Endpoint, Entity 2 and the nouns that do not belong to either of these classes)</td>
</tr>
<tr>
<td>MC₃ (Multi-class classifier, 3 classes)</td>
<td>Support vector machine multi-class classification with no additional features (3 classes, Entity 1, Endpoint, and the rest of the nouns that do not belong to either of these classes)</td>
</tr>
<tr>
<td>BC + A &amp; G (Binary classifier with Amount and Group added features)</td>
<td>Support vector machine binary classification method, one versus all + additional features, Amount and Group (OVA + Amount and Group).</td>
</tr>
<tr>
<td>MC₄ + A &amp; G (Multi-class classifier with Amount and Group added features)</td>
<td>Support vector machine multi-class classification + additional features, Amount and Group (4 classes, Entity 1, Endpoint, Entity 2 and the nouns that do not belong to either of these classes)</td>
</tr>
<tr>
<td>MC₅ + A &amp; G (Multi-class classifier with Amount and Group added features)</td>
<td>Support vector machine multi-class classification + additional features, Amount and Group (3 classes, Entity 1, Endpoint, and the noun that do not belong to either of these classes)</td>
</tr>
</tbody>
</table>

Earlier work reported results for Support Vector Machine algorithm and linear kernel18. We were interested in contrasting linear to Gaussian kernel as well as binary to multi-class classification method. Each of the experiments in Table 1 was run with linear and then contrasted with Gaussian kernel method. Of particular interest is whether the non-linear separator and transforming data into an n-dimensional space may provide better results than the linear equation on the attributes in the data set. Two types of multi-class classification experiments were conducted: 1) prediction of 2 compared entities, endpoint and all the rest of the nouns that did not belong to either of these categories (4 classes) and 2) prediction of Entity 1, Endpoint and all the rest of the nouns that did not belong to either of these categories (3 classes). Oracle Data Miner, version 3.2 was used as the platform.

The baseline approach used the same set of lexico-syntactic features was used as in earlier study (See ref 18) although a different version of the parser was used: Stanford dependency parser, version 3.5.1. Multi-class classification methods used 26 features from the previous study for the baseline model whereas binary classification method used 26 features for Entity 1 and 21 for Endpoint and Entity 218. The features used in the earlier and this study rely on the syntactic parse of the sentence provided through Stanford dependency parser and measure the syntactic but also raw distance of each of the candidate nouns from each comparison, evidence, and change anchor in the sentence. Comparison anchors represent the phrases such as compared with and similar to, different from. A set of 65 comparison anchors is used. Evidence anchor represent verbs that indicate a finding such as demonstrate, explain, transform etc. The system uses a set of 432 evidence verbs. Change anchors represent verbs that demonstrate change, such as increase, decrease, and accelerate. The system uses a lexicon of 770 change verbs. Syntactic paths that connect the syntactic root of the sentence as well as comparison anchor to each candidate noun.
are also included as the features in the model. Finally, the classifier is provided with the information on whether the candidate noun appears as a terminal leaf in the syntactic parse of the sentence. The only new features that are added in this experiment is the information on whether the head noun of the candidate noun phrase is more likely to be identified as an Amount or as a population group.

Results

Entity 1 Prediction

For prediction of Entity 1, Table 2 indicates that the results do not improve when multi-class classifier was used and no additional features were added to the model. When four classes were predicted and no additional information was added, precision dropped 0.01 point and recall 0.30 points. When we reduced the number of classes to three (focus on Entity 1 and Endpoint only), the precision increased 0.01 point but recall dropped 0.14 points. However, adding information about whether the head noun of the candidate noun phrase is likely to be categorized as Amount or Group improved precision and recall with binary classifier and linear kernel (BC + A & G). Compared to baseline (BC), precision increased 0.05 points and recall 0.02. The combination of a multi-class classifier and Gaussian kernel plus additional features (MC4 + Amount and Group) also improved the results. Precision increased from 0.39 to 0.53 (0.14 increase) and recall from 0.47 to 0.57 (0.10 increase). However, given that baseline recall was 0.58 (BC, linear) this actually represents a drop in recall of 0.01 point.

Table 2. Entity 1 results for all test set sentences (95% confidence intervals are shown in parenthesis).

<table>
<thead>
<tr>
<th></th>
<th>Linear kernel</th>
<th>Gaussian kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>(0.35, 0.41)</td>
<td>(0.34, 0.40)</td>
</tr>
<tr>
<td>Recall</td>
<td>0.58</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>(0.55, 0.61)</td>
<td>(0.25, 0.31)</td>
</tr>
<tr>
<td>F1</td>
<td>0.46</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>(0.43, 0.49)</td>
<td>(0.29, 0.35)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>(0.70, 0.76)</td>
<td>(0.74, 0.80)</td>
</tr>
</tbody>
</table>

With shorter sentences that are not longer than 30 words and that typically have fewer candidate nouns, precision improved from 0.46 using binary linear kernel classifier (BC, linear) to 0.68 using multi-class classifier and Gaussian kernel (0.22 increase) (MC3 + A & G) while the recall dropped 0.01 point from 0.63 to 0.62 (see Table 3).

Given that a series of 12 classification tasks was conducted and given that six of them did not include Amount and Group information and six did include Amount and Group information the question of interest was whether these apparent differences in the results can be attributed to chance. To establish whether adding the Amount and Group information boosts the performance by chance, a series of matched t-test on the two contrasted groups (BC, MC4, MC3—Linear and BC, MC4, MC4—Gaussian) versus (BC + A & G, MC4 + A & G, MC3 + A & G—Linear and BC + A & G, MC4 + A & G, MC3, A & G—Gaussian kernel) for each of the reported metrics was conducted. For Entity 1 prediction, the differences for individual metric—precision, recall, F1, and accuracy—on the entire test set (939 noun phrases, 132 sentences) were statistically significant (P < .05).

Table 3. Entity 1 results for short sentences (95% confidence intervals are shown in parenthesis).
In conclusion, associating the head noun of a candidate compound noun with categories such as Amount and Group improved the precision of identifying Entity 1. Compared to the baseline method, recall did not increase and typically dropped 0.01 or 0.02 points except with binary classifier, linear kernel when additional features were used \((BC + A \& G)\). Generally, multi-class classifier with additional features (regardless of the number of classes predicted) raised precision of the classifier while the recall dropped minimally.

**Endpoint Prediction**

Table 4 indicates the results for endpoint prediction. Similar to Entity 1 classification, setting the problem as a binary or multi-class classifier does not make a difference until information about the type of head noun is added. Such an addition boosts performance with both binary and multi-class classification methods. More particularly, using multi-class classifier, linear kernel, and additional features \((MC_3 + A \& G)\) on all 132 sentences improves the precision from 0.42 to 0.56 (0.14 points) and recall from 0.64 to 0.71 (0.07 improvement). Consequently, \(F_1\) measure improves to 0.62 (0.09 improvement) and accuracy to 0.79 (0.06 improvement) (Table 4).

**Table 4.** Endpoint results for all test sentences (95% confidence interval in parentheses).

<table>
<thead>
<tr>
<th>Entity 1 (385 noun phrases, 66 sentences)</th>
<th>Entity 1 (385 noun phrases, 66 sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear kernel</td>
<td>Gaussian kernel</td>
</tr>
<tr>
<td>BC</td>
<td>MC4</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Precision</td>
<td>0.46</td>
</tr>
<tr>
<td>Recall</td>
<td>0.63</td>
</tr>
<tr>
<td>(F_1)</td>
<td>0.53</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Endpoint (939 noun phrases, 132 sentences)</th>
<th>Endpoint (939 noun phrases, 132 sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear kernel</td>
<td>Gaussian kernel</td>
</tr>
<tr>
<td>BC</td>
<td>MC4</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Precision</td>
<td>0.42</td>
</tr>
<tr>
<td>Recall</td>
<td>0.64</td>
</tr>
<tr>
<td>(F_1)</td>
<td>0.51</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.73</td>
</tr>
</tbody>
</table>
With respect to short sentences ($\leqslant$30 words), the multi-class classifier with the Gaussian kernel and additional features improved the results from 0.51 (BC) to 0.58 (MC3 + A & G) and 0.59 (MC4 + A & G). Similarly, recall improved from 0.69 (BC) to 0.74 (MC3 + A & G). Consequently, the $F_1$ measure and accuracy increased to 0.63 and 0.78 (MC3 + A & G) and 0.65 and 0.78 (MC4 + A & G) (see Table 5).

With endpoint prediction, both precision and recall increase when multi-class classifier and additional information are used and we do not see the precision-recall trade-off as with Entity 1. Both types of multi-class classifiers (MC3 and MC4) and both kernel methods, linear and Gaussian, benefit from the addition of Amount and Group features. To illustrate, compared with baseline (BC) precision of 0.42, multi-class Support Vector Machine, linear kernel classifier (3 classes) achieved precision of 0.56 (0.14 increase) while recall went from 0.64 to 0.71. With shorter sentences, multi-class (3 classes) and Gaussian kernel achieved precision of 0.58 compared to 0.51 earlier best result (0.07 increase) and recall of 0.74 compared to earlier 0.64 (0.10 increase).

Table 5. Endpoint results for short sentences (95% confidence intervals are shown in parenthesis).

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>385 noun phrases, 66 sentences</th>
<th>Endpoint</th>
<th>385 noun phrases, 66 sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear kernel</td>
<td>Gaussian kernel</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BC</td>
<td>MC4</td>
<td>MC3</td>
</tr>
<tr>
<td>Precision</td>
<td>0.51 (0.48, 0.54)</td>
<td>0.51</td>
<td>0.45 (0.42, 0.48)</td>
</tr>
<tr>
<td>Recall</td>
<td>0.69 (0.66, 0.72)</td>
<td>0.59</td>
<td>0.66 (0.63, 0.69)</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.59 (0.56, 0.62)</td>
<td>0.55</td>
<td>0.54 (0.51, 0.57)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.73 (0.70, 0.76)</td>
<td>0.73</td>
<td>0.69 (0.66, 0.72)</td>
</tr>
</tbody>
</table>

In conclusion, the endpoint prediction was similar to Entity 1, where the differences between results achieved with or without Amount and Group information could not be attributed to chance. The differences were statistically significant for each individual metric (precision, recall, $F_1$, accuracy) ($P < .05$).

**Entity 2 Prediction**

Interestingly, the identification of Entity 2 does not benefit from additional information. As Table 6 indicates, the performance of the linear kernel classifier dropped after the additional information was added with both binary and multi-class classifiers.

Earlier work18 reported the closeness of Entity 2 to comparison anchor terms such as compared with, similar to, and different from. Thus, some of the best indicators for the location of Entity 2 were accompanied with a drop in precision: 0.66 compared to 0.74. In conclusion, the addition of the new features boosted the performance for Entity 1 and Endpoint but not for Entity 2 that already boasts a high level of precision and recall. (0.74 precision and 0.80 recall on longer sentences). The implication is that shattering of the search space (multi-class classifier) for Entity 2 was not helpful and did not result in better prediction results.

With short sentences, recall also increased using Gaussian kernel and binary classifier with no additional features (BC) (0.87 compared to 0.83). However, this was accompanied with a drop in precision from 0.77 to 0.71 (Table 7). In conclusion, Entity 2 prediction does not benefit from additional information and setting up the problem as a multi-class classification did not bring any improvement over binary classifier (BC). A spike in recall was recorded with Gaussian kernel and binary classification method and no additional features.
Table 6. Entity 2 results for all test sentences (95% confidence intervals shown in parenthesis).

<table>
<thead>
<tr>
<th>Entity 2 (939 noun phrases, 132 sentences)</th>
<th>Entity 2 (939 noun phrases, 132 sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear kernel</td>
<td>Gaussian kernel</td>
</tr>
<tr>
<td>BC</td>
<td>MC₄</td>
</tr>
<tr>
<td>Precision</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>(0.71, 0.77)</td>
</tr>
<tr>
<td>Recall</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>(0.77, 0.83)</td>
</tr>
<tr>
<td>F₁</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>(0.74, 0.80)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>(0.89, 0.93)</td>
</tr>
</tbody>
</table>

Table 7. Entity 2 results for short sentences (95% confidence intervals are shown in parenthesis).

<table>
<thead>
<tr>
<th>Entity 2 (385 noun phrases, 66 sentences)</th>
<th>Entity 2 (385 noun phrases, 66 sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear kernel</td>
<td>Gaussian kernel</td>
</tr>
<tr>
<td>BC</td>
<td>MC₄</td>
</tr>
<tr>
<td>Precision</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>(0.74, 0.80)</td>
</tr>
<tr>
<td>Recall</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>(0.81, 0.85)</td>
</tr>
<tr>
<td>F₁</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>(0.77, 0.83)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.94)</td>
</tr>
</tbody>
</table>

When the results of the experiments that use the additional information were compared to the results that do not use additional information were contrasted it was not clear that these individual differences cannot be attributed to chance only—the difference between these two groups for each individual metric (precision, recall, F₁, accuracy) was not statistically significant (P > .05).

Discussion

The concept of mechanism in the sciences is useful to understand the nature of comparison sentences. Mechanisms consist of entities and activities: entities are the things that engage in activities and activities are producers of change. Entities, within the context of a comparison sentence, represent things that are compared whereas activities represent the change that has occurred between the entities. The endpoint can be seen as part of the activity process, an entity, a dependent more likely than a continuant, whose role is to communicate the change that has occurred between the entities. Seen from this perspective, compared entities and endpoints would likely belong to different semantic classes. We contend that it is the addition of a semantic class to the feature set that would help separate the first compared entity from endpoint. The question, however, is how to obtain the information about the semantic
class for compared entities and endpoints? From an ontological point of view, entities that occur in comparison sentences can be matched to an ontological class such as species or population group. For example, the Unified Medical Language System Metathesaurus semantic class Population group can be seen as helpful for their identification. Consider the following sentence:

(5) The plasma insulin concentration [Endpoint_1] at 8 weeks of age [Modifier] and the pancreatic insulin content [Endpoint_2] and the beta-cell mass [Endpoint_3] on day 8 and 8 weeks of age [Modifier] in STZ-treated rats [Entity] were severely reduced compared with those of normal rats [Entity] (P < 0.001). 14988244

In this sentence, STZ-treated rats and normal rats represent two entities that can be seen as two population groups that were compared. And yet matching these two concepts to their semantic classes using the UMLS Metathesaurus does not bring us to the Population but rather to the Animals class. Neither STZ-treated rats nor normal rats has its full match in the UMLS because they represent very specific groups of rats: rats treated with streptozotocin versus rats that were not treated. With both of these examples, however, it is the head noun—rats—that provides sufficient basis for inferring the semantic class for these two entities—Mammals—and then, through its parent relation, to Vertebræ and Animals higher up in the hierarchy. Animals, however, do not link directly to Population group in the UMLS. In the higher levels of semantic network, Animals is an Entity, the broad type used for grouping conceptual and physical entities. Population group in the UMLS Metathesaurus is defined as “an individual or individuals classified according to their sex, racial origin, religion, common place of living, financial or social status, or some other cultural or behavioral attribute” and as such does not extend to Animals. Similarly, matching endpoints to an ontological class is far from being a straightforward process. Endpoints represent the processes, mechanisms, activities that are happening on the molecular, cellular, tissue, organ, or body level and as such can span semantic classes or be comprised of several semantic classes. By their nature, endpoints represent very specific processes and are typically expressed as a compound noun. In the sentence above, plasma insulin concentration, pancreatic insulin content and beta-cell mass were identified as endpoints. Matching plasma insulin concentration to its semantic class would fall under the category of a complex match as plasma insulin would be matched to one concept and concentration to another. What complicates things further is the fact that concentration also represents the case of overmatching because it is identified as a Mental Concept but also as a Quantitative Concept. It is the surrounding context that can determine the concept that should be used for concentration which in this case is a Quantitative Concept. Pancreatic insulin matches to Neoplastic Process which is not an ideal match for pancreatic insulin content. Ideally, we would have pancreatic insulin content matched to one semantic class that would be identified as the measurement of insulin in the pancreas. It is the head noun content in this compound noun that adds this quantitative quality to pancreatic insulin and steers the meaning of the noun in the direction of measurement. The situation is somewhat better with beta-cell mass, an example of a complex match. Beta cell matches to Cell semantic type and mass to Quantitative Concept semantic class, both of which identify the parts correctly.

Most often, endpoints are specific phrases and terms that sometimes indicate the outcome measure and sometimes the property of a compared entity that experienced a change. The following sentence is used to demonstrate the level of endpoint specificity:

(6) Nonfasting plasma glucose levels [Endpoint_1] and the overall glycemic excursion (area under the curve) to a glucose load [Endpoint_2] were significantly reduced (1.6-fold; P < 0.05) in (Pro_3) GIP-treated mice [Entity] compared with controls [Entity] 16046312

The endpoints in this sentence are nonfasting plasma glucose levels and glycemic excursion to a glucose load. While plasma glucose gets matched to the semantic type Laboratory procedure, nonfasting is matched to semantic class Finding which in this case is not ideal. Within the context of a comparison sentence and the information it conveys, the modifier, nonfasting provides a very important nuance for the meaning of the entire sentence and it should be retrieved as part of the endpoint.

Previous research reported a number of endpoints related to metformin drug comparison to other interventions. That study reported the following endpoints that relate to insulin: proinsulin concentrations, insulin, insulin action, insulin concentrations, serum insulin concentrations, insulin sensitivity, % suppression by insulin. The following endpoints relate to glucose: glucose, fasting glucose, hepatic glucose production during hyperinsulinemia, glucose disposal, glucose disposal rate, serum fructosamine, glycated hemoglobin (HbA1c). These endpoints were grouped based on the main substance they were measuring, insulin or glucose, whereas in this study we grouped the endpoints based on the property that they frequently share: measurement characteristic. The question remains what kind of grouping or matching system is better for the particular and specific nature of endpoints and at what modifier
and what level we can start to draw the line. These questions require a medical specialist to intervene and assist with the process of endpoint categorization.

The problems described above fall under the categories of complex, partial matches, gapped partial matches and overmatching. Most typically, endpoints are very specific phrases and it is this level of specificity that prevents them from being matched to an ontology effectively. This study showed that the fact that many endpoints lend themselves to measurement was the feature that improved their overall identification and retrieval. And yet, not all endpoints lend themselves to measurement. Consider the following sentence:

(7) There is evidence to suggest that the somatic mutational pathway \([\text{Endpoint}_1]\) may differ between invasive \([\text{Entity}_1]\) and LMP ovarian tumours \([\text{Entity}_2]\) and invasive tumours \([\text{Entity}_1]\) are more likely than LMP \([\text{Entity}_2]\) to exhibit \(p53\) overexpression \([\text{Endpoint}_2]\).

Somatic mutational pathway is identified as one of the endpoints in this sentence. It is not clear that the concept of mechanism can extend to pathways\(^2\) but even if this is the case, this type of mechanism and the change that is indicated in the above sentence does not involve measurement of any kind, only the statement that the pathway was different. Clearly, in this case, the endpoint does not lend itself to measurement in the same way as the endpoints that comprise the head noun, such as concentration, level, degree, or mass. The second endpoint, however, \(p53\) overexpression, can be measured. This sentence provides an example where one of the endpoints lends itself to measurement and the other does not indicating that the endpoints, even within the context of the same sentence, do not need to share the same characteristics.

Word ambiguity, context of the article, precision of grammatical and semantic parsers are standing in the way of better alignment of the free form of textual information in scholarly articles and with entries in ontologies and their definitions. Commonly, a large number of pre-processing tasks is needed in order to convert the text of scholarly articles to a format in which it can be matched to an ontology to enable semantic processing of the text. This study demonstrated that the identification of crucial facets of comparison sentences has benefited from additional information about the meaning of the candidate noun. The method outlined in this study requires testing on a larger dataset. Also, this work invites the examination of which semantic classes in the UMLS, or the subsets of them, can be used to indicate a Population Group effectively in the biomedical scholarly articles. Future work will also strive to examine the role of the Quantitative Concept UMLS semantic class in assisting with the process of identification, retrieval and definition of endpoints. Finally, given that not all endpoints lend themselves to measurement (for example, pathway) future work will need to establish other possible ways of modeling endpoints and establishing their significant properties that that can enable their more effective identification and retrieval.

Conclusion

The results from this study suggest that establishing if the head noun is an amount or measure enables the Support Vector Machine to differentiate nouns that play an endpoint role from other candidate noun phrases in a comparison sentence more effectively. Thus treating endpoints as an activity and seeing them through the lens of measurement provided a boost in performance with reference to their identification and retrieval. Classification performance on the test set improved for both entity 1 and the endpoint roles when amount and measure were provided as features and the improvement was statistically significant. The results were not statistically significant for Entity 2 prediction. This improved accuracy provides authors of a systematic review with more specific information about how treatments are compared. In addition, improved retrieval of endpoints will allow us to better examine and assess endpoints in aggregate and track changes over time.

References

2. Blake C, Pratt W, Tengs T, editors. Automated Information Extraction and Analysis for Information Synthesis. American Medical Informatics Association Fall Symposium (AMIA); 2002; San Antonio, TX.
5. Matwin S, Kouznetsov A, Inkpen D, Fruinza O, O'Brienis P. A new algorithm for reducing the workload of experts...


Controlling testing volume for respiratory viruses using machine learning and text mining

Mark V. Mai, MD, MHS¹, Michael Krauthammer, MD, PhD²
¹The Children’s Hospital of Philadelphia, Philadelphia, PA; ²Yale School of Medicine, New Haven, CT

Abstract

Viral testing for pediatric inpatients with respiratory symptoms is common, with considerable associated charges. In an attempt to reduce testing volumes, we studied whether data available at the time of admission could aid in identifying children with low likelihood of having a particular viral origin of their symptoms, and thus safely forgo broad viral testing. We collected clinical data for 1,685 pediatric inpatients receiving respiratory virus testing from 2010-2012. Machine-learning on the data allowed us to construct pre-test models predicting whether a patient would test positive for a particular virus. Text mining improved the predictions for one viral test. Cost-sensitive models optimized for test sensitivity showed reasonable test specificities and an ability to reduce test volume by up to 46% for single viral tests. We conclude that diverse forms of data in the electronic medical record can be used productively to build models that help physicians reduce testing volumes.

Introduction

Upper respiratory infections comprise one of the most common emergency department (ED) diagnoses in the pediatric population accounting for up to 25% of visits during influenza seasons¹². In the majority of children, the respiratory viral infection is mild and considered self-limiting with adequate supportive care³–⁵. A large variety of viruses may cause respiratory symptoms and routine testing for key viruses is now available for children visiting the ED⁶–¹¹.

The clinical utility of routine virus testing has been demonstrated previously, with children testing positive receiving less antibiotics and incurring fewer overall charges¹²,¹³. Also, children with positive results that are admitted after their ED visit may be properly isolated to prevent the spread of hospital-acquired infections¹⁴. However, with most children testing negative for the majority of viruses on the routine panels, the high billable charges of these panels, and the self-limiting nature of most virus infections, there are open questions about the overall utility of routine viral testing for ED visits.

There has been growing interest in the secondary use of EMR data to aid in clinical decision making and test ordering. Recognizing the prevalence of overtesting in the ICU setting, one group used fuzzy modeling of readily available data, like vital signs, urine collection, and transfused products, to identify tests that likely would not contribute to information gain in patients with gastrointestinal bleeding¹⁵. Other studies have looked to clinical documentation as a source of variables for developing classifiers that screen for disease. For example, emergency department free-text reports were used as the source of data for machine learning classifiers that outperformed expert Bayesian classifiers in the detection of influenza¹⁶. Another study has also employed the use of natural language processing of clinical notes to identify patients who may have Kawasaki disease, a diagnosis that is notoriously difficult to make¹⁷.

Our goal was to develop a practice-based workflow for the evaluation and optimization of respiratory virus test utilization in the pediatric population. The workflow would use data that is available at the time of the ED visit. The workflow would be helpful in identifying children with low likelihood of having a particular viral origin of their symptoms, who could safely forgo broad viral testing. The overall goal would be the design of smaller, custom viral test panels that test only those viruses that are most likely present given the pre-test patient data.

Using a retrospective study design, we collected clinical, seasonal, and demographic data for patients that underwent viral testing, and identified key characteristics that are indicative for patients with subsequent negative test results. Using machine learning, we thus optimized a classifier for the robust identification of patients with negative test results based on pre-test data. Based on current clinical knowledge of pediatric respiratory virus infections, we hypothesized that age and season would be the strongest predictors of test result when used in machine learning.
classifiers. We also hypothesized that the inclusion of clinical concepts in admission notes as predictive variables would enhance classifier performance.

Methods

The study population consisted of general pediatric inpatients who received respiratory virus laboratory testing between March 2010 and March 2012. The following inclusion criteria were employed: adequate nasopharyngeal swab sample, 18 years of age and younger, general inpatient admission, and test ordering within two days of admission. The following data were collected and de-identified for each patient: age, gender, season of test ordering, test results, and history and physical admission notes. Patients were tested with DFA and/or PCR panels. The DFA panel consisted of tests for adenovirus, influenza A/B, parainfluenza 1-3, and RSV. The PCR panel tested for the same viruses in addition to human metapneumovirus and rhinovirus. The study was performed using protocols reviewed and approved by the Yale University Institutional Review Board.

Of the clinical natural language processing (NLP) tools available, MetaMap 2013 was used to identify medical concepts in the History of Present Illness of each note. MetaMap 2013 is a program made available by the National Library of Medicine that allows users to map biomedical text to the Unified Medical Language System Metathesaurus\textsuperscript{18,19}. While clinical NLP frameworks exist that are arguably more powerful, like cTAKES, HITEX, and Sophia, MetaMap was chosen for its ease of implementation. The following options were used in MetaMap: restriction of terms to the SNOMED CT terminology, limitation of semantic types to “signs and symptoms” and “diseases and syndromes”, utilization of all derivational variants of a word, allowance for concept gaps, expansion of acronyms and abbreviations, and identification of negated concepts. The presence or absence of each mapped concept was added as additional variables for model development. To determine the precision of the MetaMap software when run on the corpus of clinical notes, concepts identified by MetaMap were compared to the original text. For 100 randomly selected notes, the identified concepts were reviewed and compared to the original statements by the physicians. Partial-match precision was calculated, as previously published\textsuperscript{20}. Recall was not calculated because we were primarily interested in the concepts that MetaMap was actually able to identify and not terms outside of its matching capabilities.

Based on these variables, we developed models using machine learning algorithms to predict the likelihood that a given patient would test positive or negative for a specific virus. For each viral diagnosis, three models were built based on: 1) billing data alone, 2) MetaMap concepts, and 3) billing data in addition to MetaMap concepts. For this task, we used Weka 3.7.9, a freely available Java based implementation that implements numerous machine learning algorithms, as well as common tools for data mining, such as data pre-processing and attribute selection\textsuperscript{21}. Using Weka, we performed attribute selection on the variables to determine which variables provided the greatest gain of information, given the specific viral test to be modeled\textsuperscript{22,23}. Attribute selection reduced the number of variables for inclusion in each model from over 400 to less than 20, depending on the test. A cost sensitive classifier was applied on top of machine learning algorithms to weigh against false negative model predictions with varying cost thresholds\textsuperscript{24}. The decision tree learning algorithm, J48 (the Java implementation of C4.5), was chosen for use in the development and evaluation of each model\textsuperscript{25}. To reduce the likelihood of overfitting, we employed ten-fold cross-validation. Hyperparameter optimization was not explored in this study.

Using the sensitivities and specificities resulting from the numerous cost thresholds, we constructed a specialized receiver operating characteristic (ROC) curve, also referred to as an ROC instance-varying transformation\textsuperscript{26}. We calculated the area under the curve (AUC) for each model using the trapezoidal rule. The standard error for each model’s ROC was calculated, as previously published\textsuperscript{27}. To simulate a scenario in which tests can be selected individually for a customized panel (e.g. in house PCR vs. multiplex PCR), we created three different types of customized panels and modeled the performance of the panels on the study sample. For each type of customized panel, a given patient would be serially tested until either a test returned positive or the patient had received all of the recommended testing. We created a control panel in which patients were assigned a panel with a random order. A second panel was composed of a fixed order based on the prevalence of diagnoses. The third and experimental panel took into consideration the predictions generated by the decision tree models. Patients would be tested for viruses, based on the recommendation of the model and then ordered from highest to lowest prevalence, and testing would cease after a positive result or all negative results for the recommended tests.

1911
Results

In total, 1,972 orders for respiratory virus testing were placed during 1,848 visits. For each patient, the test consisted of a panel of 6 viruses - adenovirus, influenza, parainfluenza, RSV, hMPV, and rhinovirus. Each hospital visit was associated with 1 or more orders for respiratory virus testing. Table 1 summarizes the age, gender, and season of the study population by etiology. Negative test results accounted for 69.5% of all tests ordered during the study period. Males comprised around half of each viral diagnosis except for adenovirus diagnoses of which males comprised 68.8%. The mean age for all viral diagnoses was less than 5 years of age, except for positive cases of influenza, where the mean age of diagnosis was 8.24.

Table 1. Summary basic clinical variables of general admission pediatric inpatients by etiology (2010-2012)

<table>
<thead>
<tr>
<th>Cases</th>
<th>Adenovirus (n = 32)</th>
<th>Influenza (n = 40)</th>
<th>Parainfluenza (n = 93)</th>
<th>RSV (n = 234)</th>
<th>hMPV (n = 29)</th>
<th>Rhinovirus (n = 180)</th>
<th>Multiple viruses (n = 57)</th>
<th>Negative (n = 1519)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males, No. (%)</td>
<td>22 (68.8)</td>
<td>21 (52.5)</td>
<td>47 (50.5)</td>
<td>125 (53.4)</td>
<td>15 (51.7)</td>
<td>91 (50.6)</td>
<td>31 (54.4)</td>
<td>858 (56.5)</td>
</tr>
<tr>
<td>Age, mean ± SD</td>
<td>3.75 ± 3.78</td>
<td>8.24 ± 6.53</td>
<td>2.99 ± 4.10</td>
<td>2.14 ± 3.20</td>
<td>4.89 ± 5.00</td>
<td>4.40 ± 4.74</td>
<td>2.39 ± 3.39</td>
<td>4.88 ± 5.29</td>
</tr>
</tbody>
</table>

Both DFA and PCR panels are ordered throughout the year and show similar patterns in peak months of test ordering, during the winter seasons. DFA panels (n = 2152) were ordered 3.9 times more often than PCR panels (n = 550). Figure 1 shows the distribution of positive test results aggregated by month during the study duration. Adenovirus and rhinovirus are detected in all months with no clear pattern. Cases of co-infection, influenza, hMPV, parainfluenza, and RSV demonstrated regular seasonal fluctuations. The distribution of co-infected cases were present at low rates during the entire year, although demonstrated peaks in the winter months of each year. No cases of influenza occurred outside the winter or spring months in the study population. Test results for hMPV had a similar distribution to influenza infections. Cases of RSV arose primarily during the winter months, although the onset of the RSV season differed between the two years.
In order to reduce viral testing in patients, we evaluated models for their predictive ability, based on pre-test data, to identify patients with low likelihood of a particular virus. Some models consumed data from text mining clinical notes. We evaluated the performance of the MetaMap program in identifying concepts by calculating the partial-match precision. The partial-match precision was calculated to be 0.724 across a random sample of 100 notes.

Figure 1. Positive test results as a proportion of all test results for each detectable virus, by month.
Using Weka, models based on the decision tree algorithm, J48, were derived. By varying the type of data used to derive the models, three types of models were developed for the detection of each virus. A basic model was derived using age, gender, and season. A text-mining model was derived using MetaMap concepts as predictive variables. A combined model was derived using age, gender, season, and MetaMap concepts. The ROC curves are depicted in Figure 2 and the corresponding AUC values are presented in Table 2. Basic models showed some predictive ability (AUC > 0.65) for four of the six viruses under study: influenza, parainfluenza, RSV, and hMPV. The text-mining models showed no predictive ability, except for RSV, where the text-mining model performed similarly to the basic model (AUC: 0.661 vs. 0.658, respectively). The combined models showed similar discriminative ability to the basic models for all viruses, except for RSV, where the combined model outperformed both the basic and text-mining.

Table 2. Receiver operator characteristic curve characteristics

<table>
<thead>
<tr>
<th></th>
<th>ROC curve AUC (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic model</td>
</tr>
<tr>
<td>Adenovirus</td>
<td>0.568 (0.114)</td>
</tr>
<tr>
<td>Influenza</td>
<td>0.743 (0.126)</td>
</tr>
<tr>
<td>Parainfluenza</td>
<td>0.686 (0.078)</td>
</tr>
<tr>
<td>RSV</td>
<td>0.658 (0.048)</td>
</tr>
<tr>
<td>hMPV</td>
<td>0.713 (0.143)</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>0.549 (0.047)</td>
</tr>
</tbody>
</table>

To determine how the MetaMap terms factor into the combined model for RSV, we mapped out the decision tree of the combined model (Figure 3). Age was the first branching point followed by branching points based on a combination of season, MetaMap concepts, and other age ranges.
Figure 3. The graphical representation of the J48 decision tree classifier for RSV when based on both billing data and MetaMap concepts. Thresholds for age and season were determined by the J48 algorithm based on the information gain provided by the attribute splits. These are denoted by the single lines. The presence or absence of clinical terms in the HPIs are shown as branching nodes, followed double lines and dashed lines, respectively.

To assess how the models would perform as “in silico” screening tests prior to actual diagnostic testing, we computed the test characteristics for the combined models. Unlike the general model presented above, we would assess models with specific performance thresholds. At a target sensitivity of ≥ 95% for the combined models for each virus, we calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) (Table 3). The target of ≥ 95% sensitivity (i.e. no more than 5% of cases with a respiratory infection are missed) was achieved by using a cost sensitive classifier that was applied on top of machine learning algorithms to weigh against false negative predictions. As detailed in Table 3, specificity ranged from 3.7% to 45.5% for all of the models at a sensitivity ~95%. A cost-sensitive classifier was thus able to safely sort patients into those that will eventually test positive for a virus (with at least 95% sensitivity), while robustly identifying patients that will not test positive (no viral infection). Some models were able to identify up to 45.5% of patients with no viral infection, and thus would lead to a sizeable reduction in test volume, as no testing is warranted. Models for adenovirus and rhinovirus showed the lowest specificities. While the PPV ranged from 50.1% to 64.7% for all models, the NPV ranged from 52.1% to 100%. For influenza, parainfluenza, RSV, and hMPV, the NPV was above 80.7%.

Table 3. Test characteristics of combined model for each of the viruses

<table>
<thead>
<tr>
<th>Virus</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>0.966</td>
<td>0.037</td>
<td>0.501</td>
<td>0.521</td>
</tr>
<tr>
<td>Influenza</td>
<td>1.00</td>
<td>0.455</td>
<td>0.647</td>
<td>1.00</td>
</tr>
<tr>
<td>Parainfluenza</td>
<td>0.965</td>
<td>0.198</td>
<td>0.546</td>
<td>0.850</td>
</tr>
<tr>
<td>RSV</td>
<td>0.953</td>
<td>0.196</td>
<td>0.542</td>
<td>0.807</td>
</tr>
<tr>
<td>hMPV</td>
<td>0.974</td>
<td>0.205</td>
<td>0.551</td>
<td>0.887</td>
</tr>
<tr>
<td>Rhinovirus</td>
<td>0.943</td>
<td>0.076</td>
<td>0.505</td>
<td>0.571</td>
</tr>
</tbody>
</table>
A minimum threshold sensitivity of 95% was set for each of the combined models (sensitivity of Rhinovirus was slightly below the set threshold).

In order to assess how model predictions might be useful in clinical practice, where patients are routinely tested for all 6 viruses, we simulated the effects of “customized panels” on the average number of tests that a given patient in the study population would need to undergo. In an environment where tests for individual viruses could be serially performed without being prohibitive in regard to time, we developed three types of customized panels that might be employed, as described in the methods section: 1) random order, 2) order based on the prevalence of a virus, 3) order based on model predictions. Each panel would test a patient serially until either a positive test result was returned or all tests on the panel had been performed. For the model based on random order, an average of 4.93 tests (SD = 1.69) would be performed; for the model based on prevalence, 4.52 tests (SD = 2.03; p < 0.001); and for the model based on the predictive models, 3.93 tests (SD = 1.64; p < 0.001). For the customized panel based on the predictive models, 22 patients would have been untested, who would have otherwise tested positive for a virus and 57 patients would have been missed for an additional viral co-infection.

Discussion

This retrospective study evaluated respiratory virus testing in the general pediatric population at a tertiary academic children's hospital. Of the tests that were ordered, the vast majority resulted in a negative finding. Our objective was to develop clinical models using information available at the time of test ordering that might reduce test volume, while ensuring that patients with detectable infections are still tested, given the previously published benefits. This study found that predictive models built on clinical variables were able to discriminate patients who should receive testing better than chance. For 4 out of 6 viruses included in our study, age, season, and gender could be used to build models with fair predictive ability. We also hypothesized that concepts contained in the history of present illness portion of the clinician’s admission documentation could improve the predictive ability of our models. In the case of RSV, a model based only on concepts in the HPI had the same predictive power as the model based on billing data. Combining these two sources of data improved prediction of RSV. The precision of our concept matches via MetaMap were in line with previously published results. Our findings suggest a useful role for admission notes, aside from documentation.

Effective implementation of these predictive models as screening tests requires practical understanding of the test characteristics as they relate to an institution’s clinical goals. By limiting testing of patients who would otherwise test negative, the population receiving diagnostic testing would be enriched for positive cases. A practical, cost-saving clinical model would thus strive for the highest possible sensitivity that is associated with a reasonable specificity. That is, high test sensitivity (almost no patient missed that will test positive) may be achieved while still achieving above zero specificity (patients with low likelihood of testing positive). We show, for example, that for influenza testing, we could achieve 100% sensitivity at 46% specificity. In other words, we can build a pre-test model that safely identifies all patients that are going to test positive for influenza, while identifying many patients that will not test positive, and for which no testing needs to be carried out. The model, as seen in Figure 1, is mostly determined by assessing the season of the patient visit.

It is important to note that the presented predictive models are not designed to predict positive cases with high accuracy. While we sought models that resulted in the highest sensitivities, these models often had very high false positive rates. Thus, in our models, the predictive model should be considered to predict “high risk for positive result” and “low risk for positive test result.” The distinction is important to the proposed function of the models, as the models were intended to aid the practitioners in reducing the number of overall tests that they order.

From a practical point of view, predictive models could act as “in silico” screening tests for whether or not patients should receive testing. If the models are validated against a prospective dataset, they might find practical application via integration into the electronic medical record, running in the background as the clinician is entering data about a patient. At the time of test ordering, the physician would be presented with information regarding what tests are likely negative or potentially positive. As a clinical decision tool at the point of care, the results of this model might encourage providers to limit testing to only those cases where the results would change clinical management. The use of basic variables, as well as documented clinical symptoms, can reduce test volume for certain tests by up to nearly half with a low false negative rate. In addition, the use of customized panels based on
the predictive models would further reduce test volume. While the actual cost of testing is not transparent at most institutions, the amount charged for performing a respiratory virus panel is estimated to be around $1500, although this varies by institution and the exact type of testing performed. Ordering clinicians and patients could therefore expect the average charge to drop to $1000 if customized panels were employed. Future work will focus on refining the text mining approach and concept identification, as well as exploring the concept of customized panels, based on the prediction results of the models.

Our study has several limitations. The study population included only patients who received respiratory virus testing and selection criteria were not based on diagnostic codes. As a result, the seasonal prevalence of test results may not be reflective the true prevalence of viral infections. Furthermore, the focus on pediatric patients only accounted for about 20% of total respiratory virus testing volume, whereas testing on adults comprises the majority of testing. In addition, the numerous exclusion criteria may limit the generalizability of the models. While the billing data are true independent variables, the variables collected by text mining with MetaMap may not be fully independent. Because the timing of when admission notes were written could not be controlled, there is the possibility that some notes may have been written after test results were communicated. We took several measures to prevent this, including selecting the oldest admission note on record and manually reviewing notes for mention of testing. The text mining approach recorded concepts as “mentioned” versus “unmentioned,” which may not fully capture the whole clinical picture for a patient, as pertinent positives or negatives may have been unknowingly omitted by the author of the HPI. Our study also did not differentiate between different strains of viruses, which some of the tests are able to do (i.e. influenza A and B, parainfluenza 1-3). Initially, we expected that concepts from clinical notes would provide more robust prediction performance than we encountered. The reasons are likely multifactorial. First, the sample size that we used was small, compared to other NLP and machine learning efforts, making it less likely that we would be able to find concepts that would be able to differentiate between viruses. Second, given the high variability between providers’ styles when writing notes, the values for many concepts were left blank and were not subsequently imputed. Finally, the clinical manifestations of upper respiratory viruses are very similar with few distinguishing symptoms (i.e. barking cough for parainfluenza, gastrointestinal symptoms with adenovirus, and myalgia for influenza), making it difficult for physicians or the program to accurately diagnose viruses based on clinical history and exam alone, thus giving value to respiratory virus testing as a means of making a final diagnosis.

Conclusion

The results presented here offer a new perspective on analyzing testing practices for respiratory viruses using data mining and natural language processing techniques. We find that in a tertiary academic children’s hospital, the majority of respiratory virus testing returns negative. Furthermore, our results suggest that additional clinical factors may be used in a clinical model to predict the likelihood of an infection and the need for further diagnostic testing. Text mining of clinical notes may augment the predictive power of future models, as demonstrated in our models of RSV. This work contributes to the growing body of evidence that diverse forms data in the electronic medical record, not just billing data, can be used productively to build models that aid physicians in decision making, and may help in reducing test volumes.

References

7. Doan Q, Enarson P, Kissoon N, Klassen TP, Johnson DW. Rapid viral diagnosis for acute febrile respiratory
10. Landry ML, Ferguson D. SimulFluor respiratory screen for rapid detection of multiple respiratory viruses in
cost-effective multiplex reverse transcriptase-PCR assay for the detection of seven common respiratory viruses in
12. Barenfänger J, Drake C, Leon N, Mueller T, Troutt T. Clinical and financial benefits of rapid detection of
emergency department: benefits for infection control and bed management. J Hosp Infect. 2011 Mar
15;77(3):248–51.
16. López Pineda A, Ye Y, Visveswaran S, Cooper GF, Wagner MM, Tsui FR. Comparison of machine learning
classifiers for influenza detection from emergency department free-text reports. J Biomed Inform. 2015
17;58:60–9.
Tool to Identify Patients with High Clinical Suspicion for Kawasaki Disease from Emergency Department
Notes. Acad Emerg Med [Internet]. 2016 Jan 30; Available from: http://dx.doi.org/10.1111/acem.12925
Inform Assoc. jamia.bmjjournals.com; 2010 May 6;17(3):229–36.
19. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc
20. Pratt W, Yetisgen-Yildiz M. A study of biomedical concept identification: MetaMap vs. people. AMIA Annu
Symp Proc. 2003 Jan 20;529–33.
23. Azhagusundari B, Thanamani A. Feature selection based on information gain. International Journal of
Inc.; 1993.
27. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve.
28. Bonner AB, Monroe KW, Talley LI, Klasner AE, Kimberlin DW. Impact of the rapid diagnosis of influenza on
physician decision-making and patient management in the pediatric emergency department: results of a
29. Chapin K. Multiplex PCR for detection of respiratory viruses: can the laboratory performing a respiratory viral
30. Dundas NE, Ziadie MS, Revell PA, Brock E, Mitui M, Leos NK, et al. A Lean Laboratory: Operational
31. Fendrick AM, Monto AS, Nightengale B, Sarnes M. The economic burden of non-influenza-related viral
Literature-Based Discovery of Confounding in Observational Clinical Data
Scott A. Malec MLIS, MSIT\textsuperscript{1}, Peng Wei PhD\textsuperscript{2}, Hua Xu PhD\textsuperscript{1},
Elmer V. Bernstam MD, MSE\textsuperscript{1,3}, Sahiti Myneni PhD\textsuperscript{1}, Trevor Cohen MBChB, PhD\textsuperscript{1}
The University of Texas Health Science Center at Houston
\textsuperscript{1}School of Biomedical Informatics; \textsuperscript{2}School of Public Health;
\textsuperscript{3}Division of General Internal Medicine, Medical School, Houston, TX

Abstract

Observational data recorded in the Electronic Health Record (EHR) can help us better understand the effects of therapeutic agents in routine clinical practice. As such data were not collected for research purposes, their reuse for research must compensate for additional information that may bias analyses and lead to faulty conclusions. Confounding is present when factors aside from the given predictor(s) affect the response of interest. However, these additional factors may not be known at the outset. In this paper, we present a scalable literature-based confounding variable discovery method for biomedical research applications with pharmacovigilance as our use case. We hypothesized that statistical models, adjusted with literature-derived confounders, will more accurately identify causative drug-adverse drug event (ADE) relationships. We evaluated our method with a curated reference standard, and found a pattern of improved performance \( \sim 5\% \) in two out of three models for gastrointestinal bleeding (pre-adjusted Area Under Curve \( \geq 0.6 \)).

Introduction

Confounding is present when factors aside from given predictor(s) affect the response of interest. For example, one may wish to understand the association between cigarette smoking and the risk of developing cancer while asbestos exposure or genetic predispositions may be present\textsuperscript{1}. Stuart Mill noted a related disparity between laboratory conditions and raw observation in nature in his System of Logic (1943)\textsuperscript{2}. Since confounding variables may not be known at the outset, we propose a scalable and computationally inexpensive method for confounding variable discovery (CVD). In this paper, we apply our method to pharmacovigilance (PV). However, confounding is a challenge to multiple fields including epidemiology, biosurveillance and pharmacogenomics.

Pharmacovigilance and FAERS. Some 770,000 adverse drug events (ADEs) occur annually in the United States alone, resulting in morbidity, mortality, and increased cost\textsuperscript{3,4}. PV aims to address the set of challenges posed by ADEs, including those detected after drugs are released to market. Recognizing the need to systematically monitor adverse effects of drugs, regulatory agencies, such as the US Food and Drug Administration (FDA), have implemented spontaneous reporting systems, through which physicians and administrators of clinical trials can report potential adverse events as they are observed. However, spontaneous reporting systems (e.g., FAERS\textsuperscript{5}) have limitations, such as incomplete clinical information, under-reporting of side-effects, and unacknowledged sources of bias\textsuperscript{3}.

Confounding in EHR. Clinical text (notes) recorded in EHRs are another potential primary source of ADE data, yet drawing reliable conclusions from routinely collected clinical data is notoriously difficult\textsuperscript{6,7}. In the PV literature, the term “signal” refers to data that are pertinent to the therapies or outcomes under study. Unlike the case of clinical trials where subjects are deliberately monitored for side-effects, clinical data were not collected for the purpose of pharmacovigilance, and are often beset with redundancy or missing data, use of non-standard abbreviations, misspellings, and so on. In addition, clinical data contain confounding variables\textsuperscript{8}. By developing more powerful methods to identify and adjust for confounding variables, we should be able to discriminate better between drug-ADE signal and noise. This in turn would facilitate the timely compilation of more comprehensive pharmaceutical risk profiles and improve public safety.
Background

Definitions. A confounding variable influences or biases the magnitude of correlation between a predictor variable (e.g., drug exposure/treatment) and a response variable (i.e. outcome/ADE). In the context of creating models for PV, when a confounding relationship exists between a falsely associated drug-ADE pair and adjustments are made to account for its influence, the association strength for that relationship, should be diminished. For example, given a set of observational clinical notes, it is observed that fish oil intake is highly correlated with acute liver injury (ALI). However, after adjusting the model for the presence of known causal agents of ALI, (e.g., acetaminophen, liver cirrhosis, hepatitis c), that correlation should approach zero. Let us consider another example of an acetaminophen exposure (predictor) and a hepatitis B infection (predictor) where the patient subsequently suffers ALI (response). In this case, each of these predictors, independently, are sufficient preconditions for ALI. As they occur together, these two predictors confound each other. When the association of either predictor is adjusted in the absence of the other predictor, the association may be diminished, but not as dramatically as in the first example. Li et al. introduced a taxonomy of confounding in PV with the following categories: confounding by indication (e.g., preexisting conditions), confounding by comorbidity (e.g., diabetes), and confounding by co-medication (e.g., aspirin) \(8\).

A mediating variable (also called an intervening variable), by contrast, lies distinctly along the causal pathway between the predictor variable and the response variable themselves and may be neither necessary nor sufficient to cause an ADE by itself. Mediators may sometimes be thought of as “risk factors” or as aspects of the etiology of the ADE itself \(8,9\). Examples of mediators include bile duct obstruction for ALI or hypertension for myocardial infarction (MI). Mediators tend to be collinear with both the predictor and the response variable, which is to say they tend to be observed together. For more detailed discussion of mediation, see Pearl \(9\).

Confounding control in PV. The two most common approaches for the control of confounding are control by design (which entails the use of case control matching, for example, where patients are matched by practice type and sometimes demographic cohort) and control by analysis \(6\). The main thrust of this paper will be the latter (while still others exist, such as counterfactual intervention, they are beyond our scope) \(6,10-13\). Control by analysis often implies the application of domain knowledge and experience to identify proxy variables in order to control for confounding. Traditionally, statisticians have depended upon the knowledge of domain experts to identify relevant confounding factors. While this is likely to result in the identification of confounding variable candidates (CVCs) pertinent to an individual study, it would be financially intractable to hire the quantity and diversity of experts needed to conduct PV across large numbers of marketed drugs and each of their potential ADEs. Li et al. recently utilized extensive search for CVCs (of the “comorbidity” subtype in PV) in observational clinical data found in EHR using penalized multivariate regression methods, specifically lasso regression \(8\). Lasso regression is a variable selection technique that shrinks multivariate predictor coefficients that fall beneath a threshold down to zero \(14\). In general, the lasso produces results that are easy to interpret and parsimonious and, in this case in particular, the results of this study are very encouraging. However, Least Angle Regression (LAR), the algorithm that is most commonly used to perform lasso regression, can be computationally expensive, depending on input, being either quadratic O(n\(^2\)) or cubic O(n\(^3\)) in its computational complexity, though recent innovations such as cyclic coordinate descent, have produced improvements in this regard \(15,16\). Other innovative signal detection approaches have involved combining multiple data sources via meta-analysis, or applying omic or biochemical substantiation techniques to verify the plausibility of the causal mechanisms that may underlie putative drug-ADE associations \(8,17-19\).

The methods of CVD mentioned above are to varying degrees financially intractable or computationally expensive, and it has been argued that brute force methods miss the “causal story” in the observational data \(20\). The purpose of our current work is to assess the extent to which feasibly-sized sets of CVCs that have been observed in the literature can be used to identify concepts for the task of confounding adjustment of clinical data. In doing so, we aim to find a middle ground between human-intensive expert-guided CVD, and computationally intensive selection of such variables based on empirical data alone.
Literature-based Discovery (LBD) is an idea first developed by Don Swanson as a means of using the biomedical literature to discover therapeutically useful associations. Swanson's approach involves finding implicit relationships between concepts that suggest an as-yet undiscovered therapeutic relationship. Recent work has incorporated semantic relations extracted from the literature using natural language processing to constrain the search space of associations. For example, the pattern of semantic relationships "drug INHIBITS x; x ASSOCIATED WITH disease" may indicate a therapeutic relationship between this drug and this disease. These patterns of relationships are known as “discovery patterns.” Our work in this area leverages high-dimensional vector space representations of the concept-by-predicate-by-concept relationships concerned to facilitate rapid search and efficient inference, using an approach called Predication-based Semantic Indexing (PSI).

To recap, our operative assumption is that if we can identify plausible therapeutic relationships in a knowledge base of literature (arguably the main focus of LBD work to date), then we can also identify confounding associations, such as associations between concepts that are exogenous to the etiology of an ADE of interest. Such concepts may be predictive of entities that tend to co-occur along with drug-ADE pairs of interest in observational clinical data. We hypothesize that statistical models of drug-ADE association will more accurately identify causative drug-ADE relationships in observational clinical data after adjustments have been made for the influence of CVCs.

Methods

Derivation of CVCs from the literature. We developed an LBD-based method that automates CVD using “false discovery patterns” (FDPs), leveraging existing NLP tools and knowledge resources.

From Predications to FDPs. MedLEE was used to perform automated coding of concepts of interest in our EHR collection. MedLEE has been shown to perform accurately on clinical notes (for example recall of .77, and precision: ~.89 for the task of extracting clinical concepts, and coding them with concept unique identifiers from the Unified Medical Language System (UMLS)). The purpose of SemRep, a publicly available NLP system developed intramurally at the US National Library of Medicine (NLM), is to identify and normalize predications - relationships in which pairs of concepts are connected by predicates (or “verbs”, e.g., “TREATS”, “CAUSES”) in biomedical literature. SemRep operates with low recall, but high precision: Kilicoglu et al. reports precision of .745, recall of .640, and an F-score of .689. SemMedDB, used as our knowledge base of biomedical literature, is a publicly-available NLM product that contains the output of SemRep processing of the entirety of MEDLINE, where concepts and predicates have been recognized and normalized, and all extracted propositions can be retrieved from a MySQL database in the following form: \texttt{ARGUMENT}_1 + \texttt{PREDICATE} + \texttt{ARGUMENT}_2. This representation disallows semantic constructs such as degree, tense, narrator reliability, or verb ditransitivity. From a practical perspective, such predication constructs are computable, since the concepts and predicates are normalized, and they facilitate our method’s inferential power.

Vector Symbolic Architectures. To process SemMedDB, which is used as our biomedical knowledge repository, we apply a representational technique called Predication-based Semantic Indexing, or PSI. PSI owes its existence ultimately to a contentious debate within the cognitive and neuroscience community in the 1980s over what became known as the parallel distributed processing (PDP) or connectionist paradigm. A number of scalable approaches to the problem of efficiently representing nested compositional structures (such as those encountered in natural language) in neural networks emerged following Smolensky’s seminal work, which utilized the tensor product as a compositional operator (Holographic Reduced Representations [HRR], Binary Spatter code, Multiply Add Permute [MAP]). This is the origin of vector symbolic architectures (VSAs), described in detail elsewhere.

PSI for efficient inferences. VSA theory provides the infrastructure for PSI. In PSI, concepts are represented by randomly generated real, binary, or complex values of high dimensionality (the more dimensions that are used generally, the better the recall and precision of the model, with a tradeoff against computational efficiency). These
are called “elemental vectors”. These elemental vectors can be superposed upon each other to generate composite semantic structures called semantic vectors \(^{39-45}\).

Random vectors are an effective way to represent elemental components, since there is a high probability of mutual near-orthogonality, particularly in higher dimensions. Semantic vectors on the other hand are composed as superpositions of the “bound products” of the elemental vectors of predicate-argument pairs (as extracted from literature using SemRep). The binding operator, which varies in implementation across VSAs, is a multiplication-like operator that provides the means to encode additional information, such as the nature and context of a relationship, into the resulting vector space. Since the same predication can be encountered in multiple documents, PSI can be thought of as a distributional model of predications. Critical to PSI, semantic representations of concepts are built up as vectors from relations found in the literature. PSI facilitates rapid search for, and retrieval of, concepts that are related to one another in particular ways (i.e., through particular predicates). As such a space is distributional, concepts in which a relationship of interest occur more frequently will be retrieved first (analogous to the way in which other information retrieval systems facilitate ranked results). In the current work, PSI is used to facilitate rapid retrieval of concepts related to other concepts in specific ways. PSI can be used to retrieve the most strongly associated concepts (called “bridging terms”) across any particular predicate pathway of interest. A discussion of predicate pathways that indicate FDPs will follow. PSI and its applications are discussed in detail elsewhere \(^{25,39-41}\).

**SemRep Predications.** For the current work, we used a PSI vector space derived from the version 24_32 of SemMedDB (processed June 20th, 2014 with version 1.5 of SemRep), containing 23.9 million citations and 70.4 million semantic predications \(^{39}\). The PSI space was built using the Semantic Vectors package (version 5.9), with 48,000-dimensional binary vectors as the elemental vectors \(^{44}\). A small number of predicates were excluded, indicating negation (such as DOES_NOT_TREAT). Terms with occurrence ≥ 500,000 were excluded.

**FDPs for CVD.** A browser interface called EpiphaNet has been developed for querying PSI-based representation of SemMedDB using a query language with meta variables that specify predicate vectors, elemental vectors, semantic vectors along with binding and superposition operators \(^{40-41,44-45}\). We can query our PSI vector space model of SemMedDB to identify CVCs (the “bridging terms” mentioned above) for each drug-ADE pair with PSI queries that represent FDPs. Below in Table 1, we present the FDP queries that we used to evaluate our CVD method. We identified these FDPs while studying how expert users of the EpiphaNet interface interpret results of drug-ADE queries, when we noticed that EpiphaNet would generate reasoning pathways that suggest confounding relationships. Also, note that “DRUG TREATS x: x COEXISTS_WITH ADE” is referred to as a “double predicate” in that it is composed of two predicates that yield CVCs that link to both the drug and ADE cue terms.

<table>
<thead>
<tr>
<th>FDP</th>
<th>PV Confounding Type</th>
<th>Examples (drug: allopurinol, ADE: liver failure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“x CAUSES ADE”</td>
<td>co-medication, comorbidity</td>
<td>transplantation, embolism, cirrhosis</td>
</tr>
<tr>
<td>“x PREDISPOSES ADE”</td>
<td>co-medication, comorbidity</td>
<td>infection, sodium</td>
</tr>
<tr>
<td>“DRUG TREATS x: x COEXISTS_WITH ADE”</td>
<td>comorbidity</td>
<td>gout, kidney failure, pericarditis</td>
</tr>
</tbody>
</table>

In our evaluation, in order to exclude spurious associations (as all vectors in the space are a measurable distance apart) from our CVCs, we made use of a frequency threshold, such that only bridging terms with association strengths ≥ 2.5 standard deviations were included.

---

1923
Figure 1. This graph illustrates EpiphaNet queries in the context of our method.

Data and Analysis

Reference Set and Data Collection. To derive our data set, we used a reference set of curated drug-ADE associations that was developed by Patrick Ryan and his colleagues as a standard for evaluating PV methods. This reference set includes 399 medications/ADE pairs and 4 ADEs with both positive (drug-ADE relationships supported by the literature and other sources, including package labeling events) and negative (drug-ADE relationships without support) control groups per ADE. The four ADEs are as follows: acute kidney injury (AKI), acute liver injury (ALI), gastrointestinal bleeding (GIB), and acute myocardial infarction (AMI). These ADEs were chosen for their importance to PV space, their diagnostics, and their impact on financial and personal cost.

Figure 2. This graph illustrates our workflow.

Analysis of Observational Clinical Data. The steps that we took for evaluation of our method are as follows both in the text below and in the corresponding items in Figure 2:

1. Map and expand drug/ADE synonyms to expose data of interest in the EHR index for extraction and additional processing [1]. We used RxNorm to identify drug synonym CUIs in the UMLS. For ADEs, we mapped the reference set’s OMOP terms (several are provide for each ADE of interest) to concepts in the UMLS, using exact string match of the preferred name field.

2. Process EHR with MedLEE to represent concepts (drugs, ADEs, CVCs) as CUIs in the UMLS. This is a subset (~2.1 million documents from ~364,000 in- and outpatients in the Houston metropolitan area between 2004 and 2012) of records from the UTH BIG clinical data warehouse.

3. Build Lucene EHR index (software that efficiently tabulates the concept co-occurrence so as to facilitate rapid document retrieval).
4. Query PSI vector space for CVCs given each drug-ADE pair for CVCs and extract data from index.
5. Extract drugs, ADEs, and CVCs from the EHR index (as lists of document IDs where respective concepts are in evidence) into concept-by-document vectors.
6. Establish baseline and CVC-adjusted scores for co-occurrence of drug-ADE pairs by constructing contingency tables [6a] and by computing area under curve (AUC) from receiver operator characteristic (ROC) curves from ranked-order of coefficients from logistic regression models [6b]. In descending order of co-occurrence count between drug, ADE, and CVC concepts (with an intersection threshold of 10), construct contingency tables as input for statistical models (using forward stepwise logistic regression), and calculate aggregated performance statistics from AUCs of ROCs. Finally, compile results into tables [6c].

In the course of building models iteratively in step 6 above, when or if a model that has incorporated a new CVC statistically converges, that CVC would be incorporated into subsequent models. This process continued for each drug-ADE pair until CVCs have been exhausted for that FDP. When such models fail to converge, the offending concept is added to a list of exclusions for that pair so that it would not be included in subsequent builds and for manual investigation of interesting patterns. If none of the models for a drug-ADE pair converge using CVCs, then the score from the unadjusted, or baseline, logistic regression coefficient is used to calculate AUC.

Results and Discussion

Table 2. This table presents the baseline and adjusted AUCs from ROCs that were calculated from ranked order of drug coefficients from logistic regression models from each ADE and FDP combination. caus="x CAUSES ADE", pred="x PREDISPOSES ADE", tcoe="DRUG TREATS x; x COEXISTS WITH ADE". Counts=number of positive/negative examples. AUCs in bold indicate that an adjusted model drug coefficient is higher than baseline.

<table>
<thead>
<tr>
<th>ADE</th>
<th>FDP</th>
<th>Complete Results</th>
<th>Constrained Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Counts (+/-)</td>
<td>Baseline</td>
</tr>
<tr>
<td>AKI</td>
<td>caus</td>
<td>24 / 64</td>
<td>.584</td>
</tr>
<tr>
<td></td>
<td>pred</td>
<td>24 / 64</td>
<td>.584</td>
</tr>
<tr>
<td></td>
<td>tcoe</td>
<td>24 / 64</td>
<td>.6126</td>
</tr>
<tr>
<td>ALI</td>
<td>caus</td>
<td>81 / 37</td>
<td>.5167</td>
</tr>
<tr>
<td></td>
<td>pred</td>
<td>81 / 37</td>
<td>.5297</td>
</tr>
<tr>
<td></td>
<td>tcoe</td>
<td>81 / 37</td>
<td>.492</td>
</tr>
<tr>
<td>GIB</td>
<td>caus</td>
<td>24 / 67</td>
<td>.653</td>
</tr>
<tr>
<td></td>
<td>pred</td>
<td>24 / 67</td>
<td>.699</td>
</tr>
<tr>
<td></td>
<td>tcoe</td>
<td>24 / 67</td>
<td>.7189</td>
</tr>
<tr>
<td>AMI</td>
<td>caus</td>
<td>36 / 66</td>
<td>.5112</td>
</tr>
<tr>
<td></td>
<td>pred</td>
<td>36 / 66</td>
<td>.5196</td>
</tr>
<tr>
<td></td>
<td>tcoe</td>
<td>36 / 66</td>
<td>.5032</td>
</tr>
</tbody>
</table>
There are two groupings of result data in Table 2, labeled Complete Results and Constrained Results. Complete Results indicates that the AUCs have been calculated from the full data set without imposing any additional criteria. In the Constrained Results, the following criteria were applied to calculate performance metrics values for each field per ADE/FDP row: all logistic regression models must have converged, the count for drug instances was \( \geq 100 \), the count for intersections was \( \geq 10 \) between med-ADR pair, and that the calculations derive exclusively from cases where CVCs were included in the logistic regression models. As a result, the count of positive and negative controls for the same ADE will vary, since CVCs differ between FDPs.

**Wordclouds.** We have generated word clouds below in Figure 3. The first two word clouds represent the CVCs of ALI and GIB with the “\( x \text{ CAUSES ADE} \)” pathway. The third and fourth show the word clouds for the AMI groups using the “\( x \text{ CAUSES ADE} \)” FDP. The fifth word cloud represents the CVCs that were excluded in building the logistic GLMs for GIB using “\( \text{DRUG TREATS } x; x \text{ COEXISTS_WITH ADE} \)” FDP. The reader will notice that CVCs from single predicate FDPs (e.g., “\( x \text{ CAUSES ADE} \)” that were associated only with the ADE were of both the comorbidity and co-medication confounding subtypes, whereas CVCs from “\( \text{DRUG TREATS } x; x \text{ COEXISTS_WITH ADE} \)” are constituted exclusively by comorbidities. The fourth word cloud from the left is interesting in that the most prevalent CVCs are comorbidities and likely mediators of myocardial infarction.

![Wordclouds](image)

**Figure 3.** These word clouds represent CVCs for ADEs and FDPs with font size proportional to aggregated term frequency (across all drugs for a group [positive or negative control] of a given ADE and FDP) in the EHR index.

**Observations.** There were \( \sim 55,000 \) instances in our EHR index for each of ALI, AMI, and GIB. AKI was the outlier with only \( \sim 5,000 \) instances. In the case of AKI for the “\( x \text{ PREDISPOSES ADE} \)” pathway, no co-occurring CVCs were identified, so no adjustment could be made. In the case of ALI, one might reason that the set is heavily weighted toward the positive examples, so little gain is to be had by adjusting with CVCs.

**Analysis of Complete Results.** Our method performed the best when there was sufficient baseline support (AUC \( \geq 0.6 \)). For example, gastrointestinal bleeding had the best baseline AUCs, and there were notable gains in both Complete Results and Constrained Results for both “\( x \text{ PREDISPOSES ADE} \)” and “\( \text{DRUG TREATS } x; x \text{ COEXISTS_WITH ADE} \)” FDPs. In general, CVCs that derived from the double predicate “\( \text{DRUG TREATS } x; x \text{ COEXISTS_WITH ADE} \)” FDP were more effective than those from single predicate FDPs. Such FDPs discover concepts with associations with both predictor and response variables, what Turing award winner Judea Pearl, a seminal influence on the field of causal reasoning, refers to as “true confounders”.

1926
Analysis of Constrained Results. In the Constrained Results, for models that include CVC adjustments, performance improved in 8 of 11 cases. The “DRUG TREATS x; x COEXIST WITH ADE” FDP improved performance in 3 of 4 cases, while same can be said in 2 of 4 for “x PREDISPOSE ADE”. In only one case, AKI, was there an improvement using the “x CAUSES ADE” FDP.

Error Analysis. While using the single predicate FDPs that imply causation or risk factors, e.g., “x CAUSES ADE”, “x PREDISPOSE ADE”, respectively, make sense in terms of their ostensible causal association for adjusting negative controls, this intuition was not supported by the results of our analysis. Such FDPs uncover concepts that exist in the gray area between mediators, risk factors (Pearl’s “indirect effects” predictors), and concepts which could manifest confounding effects, e.g., smoking with respect to a positive control drug and AMI. For example, the following examples of mediation-like CVCs from this FDP were identified: stenosis, obstruction, thrombosis, and thrombus. Such concepts are suggestive of mediating concepts that relate to the causal mechanisms for AMI. Since mediators tend to be collinear with response variables, inclusion of such CVCs may be detrimental to performance.

One perplexing CVC from “DRUG TREATS x; x COEXIST WITH ADE” FDP for acute myocardial infarction is fibroid tumor. Fibroid tumors of the gastrointestinal tract are relatively rare and usually appear on the uterus. However, review of the predication database suggested that at times anti-inflammatory agents may occur in TREATS relationships with fibroid tumors, as they are used to control the pain associated with this condition. Though the “DRUG TREATS x; x COEXIST WITH ADE” FDP is intended to retrieve terms that are associated with both the drug and the ADE, the underlying implementation involves vector superposition - so though we would anticipate terms that are bilaterally connected being retrieved first, terms that are unilaterally connected may still meet the threshold. Such spurious CVCs could be eliminated by using a higher threshold of associational strength than the 2.5 standard deviation level and by making bilateral connection a prerequisite for retrieval.

Future Work. In future iterations of our research in this area, we hope to explore more FDPs, to put in place temporal feasibility constraints, and to experiment with other reference sets. We plan to incorporate data from other sources (e.g., FAERS, plausibility models), apply causal Bayesian networks instead of regression models, explore automatically inferring FDPs from strongly associated False Positives, and experiment with interactive refinement of literature-derived CVCs.

Conclusion

With the aim of surmounting the obstacle of confounding, a phenomenon which diminishes the validity of information that can be extracted from observational data, we have proposed a scalable and computationally inexpensive LBD-based CVD method. Our results show when there is sufficient support above random for an ADE, i.e., AUC ≥ .6, that statistical models that incorporate adjustments for the influence of dual-predicate FDP-derived CVCs exhibit modest (.05 AUC or higher) performance gains for the task of re-identifying drug-ADE pairs from observational clinical data. While in the current paper we have presented a use case for our method in the domain of PV, we posit that our method is of potentially broader applicability as a tool complementary to other tools for tasks that involve enhancing signal detection from observational data within biomedicine.

Acknowledgments

This work is supported by NIH/NLM Grant 1R01LM011563, the Brown Foundation, NIH/NIGMS R01 GM103859, NIH/NCATS grants UL1 TR000371 and UL1 TR001105 and NSF grant III 0964613. We would like to thank Dr. Swaroop Gantela for proofreading this manuscript.

References

5. FAERS. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects
42. Semanticvectors package on github: https://github.com/semanticvectors/semanticvectors
48. UTH BIG. https://redcap.uth.tmc.edu/cdwstats/stats-mpi.htm
49. Lucene. https://lucene.apache.org/
Adoption of Secure Messaging in a Patient Portal across Pediatric Specialties

Mary Masterman¹, Robert M. Cronin, MD, MS¹, Sharon E. Davis, MStat¹, Jared A. Shenson¹, Gretchen P. Jackson, MD, PhD¹
¹Vanderbilt University Medical Center, Nashville, TN

Abstract

Few studies have explored adoption of patient portals for pediatric patients outside primary care or disease-specific applications. We examined use of patient-provider messaging in a patient portal across pediatric specialties during the three years after implementation of pediatric portal accounts at Vanderbilt University Medical Center. We determined the number of patient-initiated message threads and clinic visits for pediatric specialties and percentage of these outpatient interactions (i.e., message threads + clinic visits) done through messaging. Generalized estimating equations measured the likelihood of message-based interaction. During the study period, pediatric families initiated 33,503 messages and participated in 318,386 clinic visits. The number of messages sent (and messaging percentage of outpatient interaction) increased each year from 2,860 (2.7%) to 18,772 (17%). Primary care received 4,368 messages (3.4% of outpatient interactions); pediatric subspecialties, 29,135 (13.0%). Rapid growth in messaging volume over time was seen in primary care and most pediatric specialties (OR>1.0; p<0.05).

Introduction

Patient portals are secure online applications that give patients convenient, 24-hour access to personal health information with an Internet connection.¹ Government regulations like the Health Information Technology for Economic and Clinical Health Act (HITECH) and Meaningful Use (MU) have helped drive the development and implementation of patient portals.²,³

Secure messaging between patients and their providers is one of the most popular features of patient portals.⁴ Research indicates clinical care is delivered through portal message exchanges—for example, patients may report new problems or request treatment.⁵,⁶ A recent systematic review suggests secure messaging can improve medication adherence and clinic attendance.⁷ Other studies report messaging may increase patient satisfaction, improve efficiency for patients and providers, ease reporting of laboratory results, reduce costs, and increase productivity.⁴,⁸

Much of the research about secure messaging through patient portals has focused on adult primary care or chronic disease settings.⁴,⁹–¹⁴ At our institution, we have demonstrated broad adoption of secure messaging across diverse clinical specialties, with the majority of use occurring in medical or surgical specialties.¹⁵,¹⁶ A systematic review of patient portal implementation in pediatrics¹⁷ highlighted that existing research has largely focused on subpopulations of children with chronic diseases or ongoing medical conditions,¹⁸–²⁰ or have been done in primary care. The review recognized an ongoing need for more quantitative research on volume of messaging, frequency of use over time, and the impact of portal use on other types of patient contact following portal adoption.¹⁷

In this study, we aimed to characterize adoption of secure messaging by pediatric specialists, patients and their families after broad deployment of a portal at a large academic medical center. We specifically focused on uptake of messaging by primary care pediatricians and pediatric subspecialists, and examined differences in messaging rates by adolescents and their parents or other caregivers.

Methods

Setting

This retrospective cohort study was conducted at Vanderbilt University Medical Center (VUMC), which includes the tertiary care Monroe Carell Jr. Children’s Hospital at Vanderbilt (MCJCHV). VUMC implemented the My Health at Vanderbilt (MHAV) patient portal in 2005, which is available to all patients seen at VUMC. MHAV provides typical patient portal functions including secure messaging, appointment scheduling, access to selected portions of the electronic health record (EHR), account and bill management, and delivery of personalized health information. MHAV accounts for pediatric patients and their families were launched in 2007 and were fully deployed across pediatric specialties by 2008.²¹ VUMC policies for pediatric patient and parent or guardian access are similar to those for portals at other major children’s hospitals.²² Parents and guardians may access their
children’s health information through MHAV until the child turns 18 years of age. MHAV allows adolescents ages 13 years or older to have parent-controlled MHAV accounts with access to functions determined by the parent.

**Study population**

We examined all outpatient interactions at VUMC for pediatric patients between January 1, 2008 and December 31, 2010, the first three years after deployment of MHAV to pediatric patients and providers. Patients were stratified by age into three groups: children (<13 years), adolescents (≥13 and <18 years), and adults (≥18 years). MHAV users consisted of either VUMC patients (over the age of 13 years); delegates, individuals whom a patient designates to access MHAV on their behalf; or surrogates, parents or guardians who have access to their child’s or adolescent’s health information through MHAV. We defined an outpatient interaction as either a traditional face-to-face clinic visit or a patient-initiated message thread (i.e., initial message and all replies) between MHAV users and pediatric providers. For this study, we defined pediatric use as accessing of the MHAV portal by pediatric patients or any of their delegates or surrogates. Clinical providers have a designated message basket in which they receive messages from patients through MHAV, as well as through inter-departmental provider-to-provider messages. Providers access their message basket through the EHR. Each clinical unit and individual providers handle messages differently; some providers choose to answer their own messages, while other baskets are triaged first by administrative assistants or allied health professionals.

**Measures**

We collected data on all traditional outpatient clinic visits to pediatric providers during the study period. For each clinic visit, we recorded the age, gender, and race of the patient, and the specialty of the clinic. Pediatric specialties were categorized as primary care, cardiology, development, endocrinology, gastroenterology, genetics, hematology/oncology, infectious disease, neonatology, nephrology, neurology, pulmonology, and rheumatology. Visits to multidisciplinary clinics involving more than one specialty were excluded from analysis.

We also recorded the number of secure message threads initiated by pediatric accounts during the study period. For each message thread, we recorded the role of the sender (i.e., self, delegate, or surrogate); age, gender, and race of the pediatric patient; and specialty of the receiving provider. Messages threads were assigned to the same 13 specialties as the clinic visits. Messages sent to multidisciplinary or administrative message baskets were excluded.

**Analysis**

We determined the total number of unique MHAV users, message threads, recipient message baskets, and outpatient visits for each month of the study period and for each pediatric subspecialty. We constructed descriptive distributions and summary statistics for patients seen in clinic visits and patients about whom messages were sent. Generalized estimating equations with logit link controlling for age, gender, race, and time in months measured the likelihood of message-based versus clinic outpatient interaction across pediatric specialties. This analysis was conducted in R version 3.0.1, and the models were fit using the geeM package. The VUMC Institutional Review Board approved this study.

**Results**

**Patient demographics**

The demographic distribution of the pediatric patients who visited clinics or were the subjects of portal messages is shown in Table 1. Message-based interactions were more common among male and Caucasian patients (p<0.05). In models adjusting for patient demographics and clinical characteristics, the odds of a message-based rather than clinic-based interaction for African American and Hispanic patients were lower than those for Caucasian patients. Over half of all messages (58%, n=19,405) were in regard to children, with 23.8% (n=7,976) regarding adolescents and 18.3% (n=3,404) regarding adults.

**Clinic visit and messaging volumes**

During the study period, pediatric patients participated in 318,386 outpatient clinic visits. The total number of visits across specialties was relatively stable over the course of the study (Figure 1A). Primary care had the most clinic visits at 122,773 visits during the study period, followed by endocrinology (37,337), hematology/oncology (34,831), and gastroenterology (22,506). Infectious disease, neonatology, and palliative care had the fewest clinic visits. 81 Palliative Care outpatient visits were excluded, as there were no message threads associated with the specialty. 3430 encounters with Allergy and Immunology and 148 messages and encounters associated with Adolescent Medicine.
were excluded as these services are multidisciplinary. For the purposes of the analysis, 47 encounters and messages associated with Heart Transplant were grouped with Cardiology.

Table 1. Demographic distributions across pediatric outpatient interactions.

<table>
<thead>
<tr>
<th></th>
<th>Outpatient Encounters (N=318,386)</th>
<th>Message Threads (N=33,503)</th>
<th>Odds Ratio of Message-based vs. Clinic-based Interaction(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (% Total Encounters)</td>
<td>N (% Total Messages)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>152,695 (48.0%)</td>
<td>15,705 (46.9%)</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>Male</td>
<td>165,691 (52.0%)</td>
<td>17,798 (53.1%)</td>
<td>1.11 (1.08, 1.14)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>69,333 (21.8%)</td>
<td>1,999 (6.0%)</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>American Indian/Eskimo</td>
<td>525 (0.2%)</td>
<td>14 (0.0%)</td>
<td>0.85 (0.48, 1.50)</td>
</tr>
</tbody>
</table>
| Asian/Pac.
Islander | 7,059 (2.2%)                     | 423 (1.3%)                | 2.14 (1.91, 2.40)                                          |
| Caucasian        | 156,358 (49.1%)                    | 28,131 (84.0%)            | 4.29 (4.08, 4.52)                                          |
| Hispanic/Latino  | 31,253 (9.8%)                     | 179 (0.5%)                | 0.21 (0.18, 0.25)                                          |
| Unknown          | 53,858 (16.9%)                     | 2,757 (8.2%)              | 1.30 (1.22, 1.39)                                          |

\(^a\)Odds ratios were computed using the generalized estimating equations model. This model controlled for age, gender, race, and time in months in order to measure the likelihood of message-based versus clinic outpatient interaction across pediatric specialties.

Figure 1. Volumes of outpatient interaction types. Number of (A) outpatient clinic encounters, (B) message threads, (C) recipient message baskets, and (D) the messaging percentage of outpatient interaction per month. Month 0 corresponds to January 2008; month 35 corresponds to December 2010.
From 2008 to 2010, pediatric patients and their families initiated 33,503 message threads, with the number of messages increasing from 2,860 in the first year to 11,871 and 18,772 in the second and third years, respectively (see Figure 1B). The number of provider baskets receiving messages increased from 346 in the first year to 701 and 1017 in the second and third years, respectively (Figure 1C). As there was only one message thread associated with Emergency Medicine, and this specialty is not associated with traditional outpatient visits, the message was excluded from the analysis. 894 messages associated with Allergy and Immunology were excluded for the reason stated above.

Users and user roles

Message threads were initiated by 19,049 unique users, including 13,848 surrogates (72.7%), 4,747 patients themselves (24.9%), and 454 delegates (2.4%), and received by 2,064 unique provider message baskets. The distribution of user types sending messages by month for children and adolescents is shown in Figure 2. Of the 10,747 users who messaged on behalf of children younger than age 13, 10,452 (97.3%) were surrogates and 261 (2.4%) were delegates. Forty-six messages (1.3%) were sent by the children themselves; in the early stages of pediatric account deployment, some children did have their own accounts created inadvertently. Of the 4,202 users who messaged on behalf of adolescents, 3,328 (79.2%) were surrogates, 99 (2.4%) were delegates, and 775 (18.4%) were the adolescents themselves. Although the number of adolescents using messaging increased from 103 in 2008 to 304 and 368 in 2009 and 2010, respectively, the percentage of total users who were adolescents decreased over time. Of the message exchanges with pediatric providers regarding adult patients, 68 (1.7%) of the 4,100 who messaged on behalf of adult accounts were surrogates, 94 (2.3%) were delegates, and 3,938 (96%) were the patients themselves.

Messaging across clinical specialties

Pediatric gastroenterology received the most messages (n=9,399; 28.1%) over the study period, followed by endocrinology (n=6,037, 18%), primary care (n=4,368; 13%), and neurology (n=3,717; 11.1%) (Table 2). Specialties receiving the fewest messages included development, infectious disease, genetics, and neonatology. With the exception of genetics, neonatology, and hematology/oncology, all pediatric specialties experienced significant growth (p-value<0.05; OR>1.0) in message volume over time during the study period when compared to primary care.

Over the study period, pediatric patients and their families used messaging for 10% of all outpatient interactions (i.e., message threads plus clinic visits). The proportion of pediatric outpatient interaction conducted through messaging rose from 2.7% in 2008 to 10.3% in 2009 and 14.6% in 2010 (Figure 1D). The proportion of message-based outpatient interaction for pediatric primary care was 3.4%, while the mean proportion of outpatient interaction across pediatric subspecialties was 19.1%. The specialties with the highest utilization of messaging as a form of outpatient interaction were gastroenterology (29.5% of outpatient interactions), pulmonology (15.5%), neurology (14.7%), and rheumatology (14%) (Table 2). The specialties with the lowest percentage of messaging-based outpatient interaction were neonatology (1.5%), infectious disease (1.2%), genetics (0.1%), and palliative care (0%).

Figure 2. Distribution of users messaging on behalf of A. children and B. adolescents. Month 0 corresponds to January 2008; month 35 corresponds to December 2010.
outpatient interactions by specialty.

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Total Outpatient Encounters (N=318,386)</th>
<th>Total Message Threads (N=33,503)</th>
<th>% of Outpatient Interactions that are Messages</th>
<th>OR of Message-based vs. Clinic-based Interaction^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Care</td>
<td>122,773 (38.6%)</td>
<td>4,368 (13.0%)</td>
<td>3.4%</td>
<td>1 [Reference]</td>
</tr>
<tr>
<td>Cardiology</td>
<td>22,280 (7.0%)</td>
<td>2,017 (6.0%)</td>
<td>8.3%</td>
<td>0.70 (0.58, 0.84)</td>
</tr>
<tr>
<td>Development</td>
<td>11,969 (3.8%)</td>
<td>574 (1.7%)</td>
<td>4.6%</td>
<td>0.26 (0.19, 0.36)</td>
</tr>
<tr>
<td>Endocrinology</td>
<td>37,337 (11.7%)</td>
<td>6,037 (18.0%)</td>
<td>13.9%</td>
<td>1.03 (0.91, 1.15)</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>22,506 (7.1%)</td>
<td>9,399 (28.1%)</td>
<td>29.5%</td>
<td>3.35 (3.01, 3.73)</td>
</tr>
<tr>
<td>Genetics</td>
<td>7,554 (2.4%)</td>
<td>10 (0.0%)</td>
<td>0.1%</td>
<td>0.00 (0.00, 0.00)</td>
</tr>
<tr>
<td>Hematology/Oncology</td>
<td>34,831 (10.9%)</td>
<td>1,371 (4.1%)</td>
<td>3.8%</td>
<td>0.53 (0.45, 0.62)</td>
</tr>
<tr>
<td>Infectious Disease</td>
<td>3,217 (1.0%)</td>
<td>39 (0.1%)</td>
<td>1.2%</td>
<td>0.04 (0.02, 0.10)</td>
</tr>
<tr>
<td>Neonatology</td>
<td>398 (0.1%)</td>
<td>6 (0.0%)</td>
<td>1.5%</td>
<td>0.03 (0.01, 0.17)</td>
</tr>
<tr>
<td>Nephrology</td>
<td>8,455 (2.7%)</td>
<td>1,195 (3.6%)</td>
<td>12.4%</td>
<td>1.30 (1.10, 1.55)</td>
</tr>
<tr>
<td>Pulmonology</td>
<td>15,185 (4.8%)</td>
<td>2,781 (8.3%)</td>
<td>15.5%</td>
<td>1.76 (1.54, 2.00)</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>6,737 (2.1%)</td>
<td>1,095 (3.3%)</td>
<td>14.0%</td>
<td>1.01 (0.80, 1.26)</td>
</tr>
</tbody>
</table>

^a Odds ratios were computed using the generalized estimating equations model. This model controlled for age, gender, race, and time in months in order to measure the likelihood of message-based versus clinic outpatient interaction across pediatric specialties.

Discussion

This study demonstrates rapid adoption of secure messaging through a patient portal by pediatric patients, families, and providers, with the total volume of messaging increasing by more than six times during the first three years after deployment of pediatric accounts. There was also substantial growth in the proportion of outpatient interaction involving messaging rather than in-person encounters, increasing from 2.7% in 2008 to 14.6% by 2010. In this study, a large number of adolescent patients exchanged messages with their providers; by 2010, 16% of all those users who were messaging on behalf of adolescents were adolescents themselves. To our knowledge, this study is the first to characterize adoption of portal messaging across a diverse set of pediatric patients and clinical specialties.

We found pediatric primary care experienced lower volumes and less growth in messaging relative to pediatric subspecialties, which differs from usage patterns that have previously been reported for other institutions, primarily in adults. Although primary care clinics handled over a third of all pediatric clinic visits, only 13% of all messages were sent to pediatric primary care providers. There are several possible reasons for this observation. First, children seeing primary care physicians tend to be healthier than children seeing subspecialists, with a large proportion of outpatient visits dedicated to well child visits which may not need a lot of follow-up or extended communications that would be well-suited to portal-based messaging. Second, acute issues seen by pediatric primary care providers may typically resolve without needing follow-up, such as viral infections. Third, the MCICHV is a tertiary care center with a large rural referral base. Some families may have visited specialist clinics for specific needs and then followed up with pediatricians outside the VUMC network. The middle Tennessee area has many private pediatrics practices not affiliated with VUMC. These providers would not have access to the MHAV patient portal and may have their own patient portal; thus, such patients and their families would be less likely to use MHAV messaging.

In contrast, pediatric specialists, such as gastroenterologists, who managed a high volume of messages over the course of the study (29.5% of outpatient interactions), frequently provide care for children who have chronic diseases, such as short gut syndrome and inflammatory bowel disease, which require long-term nutrition and symptom management. Management of such long-term care may lend itself to message-based interactions between clinic visits. Gastroenterology is also a highly procedural specialty, and patients and their families may use the portal to follow up on the results of biopsies from endoscopies and other procedures. Pediatric endocrinology, neurology, and pulmonology, which also experienced high volumes of messaging, are also specialties that involve chronic conditions (e.g., diabetes, seizure disorders, and cystic fibrosis) that may require repeated interactions between
Figure 3. Use of secure messaging across pediatric specialties. Number of office visits, number of message threads, and the percentage of outpatient encounters that are messages are shown for each specialty. Number of office visits and message threads are plotted on the left y-axis. % of outpatient encounters as messages is plotted on the right y-axis. Months are plotted on the x-axis from month 0 (January 2008) to month 35 (December 2010).
families and providers for medication or symptom management, making secure messaging a valuable tool. In contrast, the specialties of pediatric genetics, development, and neonatology, which had lower volumes of clinic visits, may involve fewer interactions with patients and family, with little ongoing interaction after establishing a diagnosis. Some variation in adoption of secure messaging across specialties also may be due to individual differences in group practices at VUMC. Barriers to adoption of secure messaging that have been described for adult populations in the past have included providers’ workload concerns, clinical communication preferences, and clinical adoption inertia resulting from the information overload in a clinical environment that is already highly computerized.\textsuperscript{25} The reasons for the differences in usage of patient portal messaging by the pediatric primary care and subspecialty populations are not evident from our usage data alone. Our ongoing research is exploring the content of portal messages across specialties, which may yield further insight into these differences. Future research using qualitative methods such as interviews or focus groups with patients and physicians may elucidate the reasons for differences in usage.
Prior work has shown a complex relationship between the implementation of messaging and effects on the number of traditional clinic visits. Some studies reported declines in the number of traditional clinic visits after messaging implementation,\textsuperscript{10,11,26} suggesting that clinical care may be offset; others have observed no change or increases in office visits after messaging is implemented.\textsuperscript{27,28} Although we were not able to obtain data about the baseline number of clinic visits, across specialties, the number of clinic visits per month did not change significantly over the study period after the portal was implemented.

There are important considerations uniquely relevant to use of patient portals and secure messaging in pediatrics. Adolescents' and parents' legal rights to access patient records are not standardized and vary between states. MHAV requires parents to register to access their child’s accounts, which enabled us to contrast usage between patients, delegates, and surrogates. We found that, as expected, most of those messaging about children were surrogates (97%), with the rest being the small number of delegates who had been given access to their record. However, adolescents sent their own messages 18% of the time. Of all messages sent during the study, 18% were sent about adult patients, an observation due to the continued care of patients over the age of 18 by pediatric specialists, usually for congenital disorders or diagnoses common during childhood. Of those who messaged pediatric providers on behalf of adults, 96% were the adults themselves, 2% were surrogates, and 2% were delegates.

This study has several limitations. First, it was conducted at a single large academic medical center with a locally developed patient portal with specific features and attributes that may limit the generalizability of the findings to other institutions.\textsuperscript{21} Second, VUMC has approached pediatric accounts like other major children’s hospitals with policies that focus on preserving parental and guardian rights over those of adolescents\textsuperscript{21,22} (i.e., the “family engagement” model). As such, parents have access to all of their child’s health data and control access rights for their children. In contrast, the University of Florida implemented an adolescent patient portal with very limited access to adolescent patient health data by parents or guardians, also known as the confidentiality model.\textsuperscript{29} This approach was designed to comply with known adolescent care best practices as well as state laws that protect confidentiality of teenage patients. In this study, although the rate of portal and messaging use by adolescent patients themselves was high, our policies that allow ongoing parental access for adolescent patients may have limited adolescent use, especially for sensitive primary care issues such as birth control and sexually transmitted infections. In addition, it’s possible that some of the adolescents or adults who sent messages themselves were actually caregivers who were logged into the patient’s account. However, VUMC discourages this practice by requiring surrogate account use for young children and then keeping the surrogate accounts separate from the patient accounts when patients turn 13.

Our study is limited in that MHAV does not store complete sociodemographic information for all delegates and surrogates, and thus, our analysis included only the characteristics of the patient about whom the message was sent. Thus, we cannot explore whether there were age- or race-based disparities in usage of the caregivers. The focus of this work was messaging use across pediatric specialties, as age and racial disparities in portal usage have been explored elsewhere.\textsuperscript{30,31} In addition, this paper describes data from 2008 to 2010, and both healthcare institutions and consumers have had substantial experience with patient portals since that time. Thus, contemporary usage patterns, which may be influenced by the Affordable Care Act and Meaningful Use incentives and penalties, may vary. We selected this time period to focus on the usage of a portal immediately after implementation of accounts for pediatric patients and their caregivers. Thus, our results may inform other institutions that are at the early stages of pediatric portal adoption. Our research group is currently examining long-term trends in usage of the portal across specialties.

In this study, we only explored use of secure messaging, not looking at other functions, such as access of health education topics or access of selected portions of the EHR. This decision was made because messages are sent to baskets associated with a certain specialty, but it would have been difficult to determine the specialties associated with use of all other functions. We also assumed that the specialty associated with the initial message within each thread was the most appropriate specialty for the entire thread of messages, not accounting for potential transfers to other clinical specialties. We did not examine message content to determine what types of care or other services were provided through these messages, although this topic is the focus of ongoing research. This study did also not look at other forms of outpatient interaction beyond messaging and clinic visits, namely, telephone calls and non-
secure emails. At VUMC, telephone calls are not systematically recorded, preventing quantification of call volumes; email interaction with patients is strongly discouraged. Finally, in our analysis, the GEE modeling approach allowed us to cluster records of message threads associated with the same patient, but we were unable to cluster encounters by patient. Therefore unaccounted correlations for interactions involving the same patient might affect the estimated standard errors. However, given the large sample size and high level of significance in the observed p-values, our findings are unlikely to be affected.

Conclusions

This study is one of the first to analyze use of secure messaging in a patient portal for pediatric patients across clinical subspecialties. After widespread deployment of pediatric accounts at our children’s hospital, most specialties experienced rapid growth in adoption of secure messaging. In contrast to studies of portal use for adult patients, we found that pediatric subspecialties demonstrated greater growth in messaging use as a form of outpatient interaction compared to primary care. Our research may inform institutions that are at the early stages of patient portal adoption or deployment of pediatric accounts. As different stages of MU unfold across the country and pediatric practices adopt patient portals and secure messaging, providers can expect to see rapid increase in the usage of messaging as a form of outpatient interaction, not only in primary care but also across pediatric subspecialties. In adopting patient portals at a children’s hospital, adolescents require unique considerations. Our study revealed enthusiastic adoption by adolescent patients across clinical specialties with use of patient-provider messaging by both adolescents and their parents. When implementing portal messaging for the pediatric population, institutions must take care not to violate state privacy laws and protect the privacy of adolescents. Additional research is needed to examine the types of information and care delivered through patient portal messages in pediatric specialties.

References


1938

Justin Mower BS¹,², Devika Subramanian PhD³, Ning Shang PhD⁴, Trevor Cohen MBChB, PhD¹,²
¹Baylor College of Medicine, Houston, Texas; ²University of Texas Health Science Center at Houston, Houston, Texas; ³Rice University, Houston, Texas; ⁴Columbia University, New York, New York

Abstract

An important aspect of post-marketing drug surveillance involves identifying potential side-effects utilizing adverse drug event (ADE) reporting systems and/or Electronic Health Records. These data are noisy, necessitating identified drug/ADE associations be manually reviewed – a human-intensive process that scales poorly with large numbers of possibly dangerous associations and rapid growth of biomedical literature. Recent work has employed Literature Based Discovery methods that exploit implicit relationships between biomedical entities within the literature to estimate the plausibility of drug/ADE connections. We extend this work by evaluating machine learning classifiers applied to high-dimensional vector representations of relationships extracted from the literature as a means to identify substantiated drug/ADE connections. Using a curated reference standard, we show applying classifiers to such representations improves performance (+≈37%AUC) over previous approaches. These trained systems reproduce outcomes of the manual literature review process used to create the reference standard, but further research is required to establish their generalizability.

Introduction

In 2007, the Institute of Medicine estimated that 1.5 million preventable adverse drug events (ADEs) occur each year in the United States¹. One report in JAMA indicated that ADEs are the most common nonsurgical adverse events that occur in hospitals,² and another meta-analysis indicated that ADEs were between the fourth and sixth leading cause of patient mortality³. Additionally, 25 drug products were removed from market due to safety issues over the last decade, highlighted by high-profile cases such as Vioxx (rofecoxib)⁴, which was removed from market on account of increased risk of potentially fatal cardiovascular side effects. The seriousness and prevalence of post-marketing ADEs is a motivation for modern pharmacovigilance systems – systems that actively monitor adverse event reports and clinical records for the emergence of yet undetected associations between drugs and side effects. A key challenge to this process, however, is determining whether there is sufficiently compelling evidence to support the belief that an observed drug/side-effect relationship is plausibly causal⁵. In order to make this assessment, information from a variety of data sources, including randomized clinical trials, observational studies, and spontaneous ADE reporting systems, is integrated by subject matter experts⁵. This process is extremely time- and resource-intensive, and doesn’t scale well to the vast and growing amount of such data. Though there has been a considerable amount of methodological research focused on the problem of signal detection – the selective identification of meaningful drug/ADE associations using statistical methods – an urgent need exists for informatics methods to support the process of critical clinical review to establish the plausibility of such associations once identified⁶.

In this paper, we evaluate a novel approach to this problem by applying machine learning methods to vector representations of implicit relationship patterns that connect a given drug/ADE pair in the literature. We call this approach "Classification-by-Analogy", as classification is thought to occur on the basis of the alignment between the relational structures connecting pairs of entities - a defining characteristic of analogical reasoning⁷. Our hypothesis is that the structure (rather than just the content) of the relations that connect pairs of entities in the literature can serve as a meaningful basis for categorization of the nature of the relationship between them. As an initial case study, we set out to determine whether a drug/ADE pair has a plausibly causal relationship substantiated in the literature, as determined by expert review.

Background

In the field of Literature-based Discovery (LBD), relationships extracted from the literature are used to establish the plausibility of an observed or hypothetical relationship (known as “closed discovery”)⁸-¹⁰. The main idea is that two concepts (such as a drug and disease) that are not connected directly in the literature, may be connected implicitly by relations involving other concepts (for example a drug might inhibit a gene associated with a disease). Though originally intended to assess potentially therapeutic relationships¹⁰, this approach can also be applied to drug/ADE
relationships\textsuperscript{11,12}. Several systems have been deployed in an effort to automate LBD analyses, many of which operate on a co-occurrence based approach\textsuperscript{13-16}. In this paradigm, co-occurrence of concepts or terms are taken as indications of relationship between concepts of interest. In general, these methods do not consider the nature of a particular relationship, even if assertions that specify it occur in the text. These linking words are of particular interest in LBD methods, however, as they can add additional information to constrain the search space of intermediate concepts. Natural language processing systems such as SemRep have been developed to extract these relational assertions from the biomedical literature\textsuperscript{17}. SemRep preserves relational assertions by extracting concept-predicate-concept triplets – such as (ibuprofen-TREATS-pain) – which have been used effectively for LBD in a number of applications\textsuperscript{15,18}. As noted by LBD’s originator, Don Swanson, however, exhaustive exploration of every implicit relationship occurring between concepts is unlikely to be computationally tractable\textsuperscript{19}, motivating the development of methods that operate on reduced-dimensional approximations of the relationship matrix between concept pairs\textsuperscript{19,20}. In one approach, concept-predicate-concept triplets form the basis for a vector representation scheme, Predication-based Semantic Indexing (PSI)\textsuperscript{19,21}, that encodes concepts and their relational connections (or predicates) in a hyper-dimensional semantic vector space. These PSI encodings can be used to query the relationships between concepts by using an approximate form of reasoning in which the potentially intractable task of exploring large numbers of implicit relationships is converted to the computationally convenient task of comparing the similarity between semantic concept vector representations\textsuperscript{21}. Although PSI represents concepts as vector encodings in a high-dimensional vector space similar to those of neural embedding approaches (e.g. word2vec\textsuperscript{22}), these encodings are derived differently. Neural word embeddings are derived directly from natural language using a neural network optimized to predict the context of a given term (or vice versa). In contrast, PSI representations are derived from semantic predications by explicitly encoding the nature of the relationships between concepts using compositional operators.

PSI accomplishes encoding and query functions by leveraging reversible vector transformations provided by a family of representational approaches: Vector Symbolic Architectures (VSAs)\textsuperscript{23-25}. The vector transformations are algebraic operations, and can be characterized as follows: a bundling operation (+), which adds (or superposes) vectors to generate a vector product that is similar to its component vectors; and binding (\(\odot\)), which results in a vector that is dissimilar from its component vectors and is functionally analogous to multiplication. These vector transformations are reversible by subtraction of component vectors and release (\(\odot\)) of the binding operation respectively, and vary in technical implementation between VSAs. In our work, we employ the Binary Spatter Code (BSC) as the VSA, which uses high- (or “hyper-”) dimensional binary vectors with dimensionality on the order of 1,000s as a representational unit\textsuperscript{26}. On account of the statistical properties of high-dimensional space, large numbers of such vectors can be generated stochastically – by randomly assigning a one or zero in each dimension with equal probability – with a high probability of their being far apart in space. This means these vectors are exceedingly unlikely to be confused with one another, despite their being distorted during the superposition process: large numbers of such random vectors can be superposed before their signal is lost. The BSC’s bundling transformation takes a majority rule vote between component vectors (ones are assigned to dimensions with more ones than zeros, and ties are broken at random), and employs Pairwise Exclusive OR (XOR) to bind and release (since XOR is its own inverse). These operations provide the basis for training in PSI. The semantic vector for a concept is generated by superposing the bound product of the random vectors for the predicate and argument of each predicate it occurs in. For example, the predication “ibuprofen-TREATS-pain” would be encoded into the semantic vector for the concept “ibuprofen” by superposing the bound product of the random vectors for “TREATS” and “pain”. In symbols, where \(S(\text{concept})\) is the semantic vector for a concept, and \(E(\text{concept/PREDICATE})\) is the elemental (or random) vector for a concept or a predicate, \(S(\text{ibuprofen}) \equiv E(\text{TREATS}) \odot E(\text{pain})\). A consequence of this encoding process is that when applied to two semantic vectors, the release operator reveals the two-predicate path (if any) that connects them:

\[
\begin{align*}
\text{If} & \quad S(\text{ibuprofen}) & \equiv & \quad E(\text{TREATS}) \odot E(\text{pain}) \\
\text{and} & \quad S(\text{arthritis}) & \equiv & \quad E(\text{CAUSES}) \odot E(\text{pain}) \\
\text{then} & \quad S(\text{ibuprofen}) \odot S(\text{arthritis}) & = & \quad E(\text{TREATS}) \odot E(\text{pain}) \odot (E(\text{CAUSES}) \odot E(\text{pain})) \\
& & = & \quad E(\text{TREATS}) \odot E(\text{pain}) \odot E(\text{pain}) \odot E(\text{CAUSES}) \\
& & = & \quad E(\text{TREATS}) \odot E(\text{CAUSES})
\end{align*}
\]

1941
For example, in the PSI semantic space for the current work, the nearest-neighboring bound products of predicate pairs to the vector product $S(\text{valdecoxib}) \otimes S(\text{gastrointestinal_hemorrhage})$ are shown in Table 1. Importantly, these pathways contain only bridging relationship information (structure), and not the bridging concepts themselves (content).

Once inferred, these predicate-based (or reasoning) pathways can be used to direct a search through the space for other concepts that relate to a third concept in a manner similar to the relationship between the cue concept pair—a process referred to as Discovery-by-Analogy ($DbA^2$), as the reasoning employed follows the pattern: “what is to myocardial infarction as valdecoxib is to gastrointestinal hemorrhage”. We have previously applied this methodology to estimate the plausibility of drug/ADE relationships, using a procedure that restricts the search to a small number of two- and three-predicate pathways (termed “discovery patterns”) inferred from known therapeutic or drug/ADE relationships.

In this paper, we take previous PSI approaches a step further: rather than inferring and applying discrete reasoning pathways from known examples (the “discovery patterns”), we evaluate the utility of applying machine-learning methods to the vector products of semantic PSI representations directly. As these vector products represent patterns of relationships (predicates), we call this approach Classification-by-Analogy ($CbA$). Our hypothesis is that this will lead to improved performance in a classification task, as the distribution of reasoning pathways between pairs of concepts, rather than the strength of relatedness across a set of discrete pathways, is considered.

By way of novelty, these models have not yet been utilized as a representational framework for machine learning in the biomedical domain, aside from in the context of $DbA$, so the literature provides little guidance as to which algorithms might be best applied to them. For the current analyses, we chose to utilize k-Nearest Neighbors ($kNN$), a support vector machine ($SVM$), and a logistic regression ($LR$) model. $kNN$ is a nonparametric classifier that functions in simple deployments by taking a majority vote amongst the closest $k$-neighbors to an unknown data point. Since VSAs generate a vector space populated by vectors in such a way that similar vectors co-localize to a similar geometric region, we anticipated that $kNN$ would provide reasonable performance. SVMs are parametric models that learn a dividing hyperplane defined by a subset of the data (so-called support vectors) in high dimensional spaces to classify examples occurring on either side. As one previous example exists of an SVM applied to vector symbolic representations with success on a text categorization task, an SVM with similar parameters was chosen. Finally, $LR$ was chosen due to both its popularity and its simplicity in defining a classifying hyperplane as a function of coefficients on the input data alone (i.e., it does not learn the hyperplane as a function of support vectors). Cost functions differ between SVMs and $LR$, and they differ slightly in their handling of regularization despite sharing the same hyperparameter, $C$, a term to encourage sparsity.

Labeled data are required input for these supervised machine learning algorithms. In pharmacovigilance, such labeled data has historically been difficult to acquire, as it requires the very human-intensive process that makes pharmacovigilance itself a costly expenditure. The Observational Medical Outcomes Partnership (OMOP) research initiative endeavored to meet this data need, and produced a drug/side-effect database to facilitate methodological research for drug safety surveillance. This manually curated reference set consists of 165 positive

<table>
<thead>
<tr>
<th>Rank (std &gt; mean)</th>
<th>Neighboring Predicate Pathway</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (4.979988)</td>
<td>$E(\text{COMPARED_WITH}) \otimes E(\text{PREDISPOSES-INV})$</td>
<td>Valdecoxib was compared with (e.g. in a clinical trial) a drug that predisposes toward gastrointestinal hemorrhage (gih).</td>
</tr>
<tr>
<td>2 (4.031143)</td>
<td>$E(\text{COMPARED_WITH}) \otimes E(\text{CAUSES-INV})$</td>
<td>Valdecoxib was compared with a drug that causes gih.</td>
</tr>
<tr>
<td>3 (2.931345)</td>
<td>$E(\text{ISA}) \otimes E(\text{CAUSES-INV})$</td>
<td>Valdecoxib is of a class of agents that causes gih.</td>
</tr>
<tr>
<td>4 (2.801957)</td>
<td>$E(\text{COMPARED_WITH}) \otimes E(\text{TREATS-INV})$</td>
<td>Valdecoxib was compared with an agent that treats gih.</td>
</tr>
<tr>
<td>5 (2.780393)</td>
<td>$E(\text{COEXISTS_WITH}) \otimes E(\text{PREDISPOSES-INV})$</td>
<td>Valdecoxib coexists with a condition that predisposes toward gih.</td>
</tr>
</tbody>
</table>

Table 1. The closest five predicate paths to $S(\text{valdecoxib}) \otimes S(\text{gastrointestinal_hemorrhage})$. -INV indicates directionality, such that CAUSES-INV can be read as “is caused by”. The rank amongst predicate pathways, and the standard deviation above the mean similarity score across these vectors are shown in the first column.

For example, in the PSI semantic space for the current work, the nearest-neighboring bound products of predicate pairs to the vector product $S(\text{valdecoxib}) \otimes S(\text{gastrointestinal_hemorrhage})$ are shown in Table 1. Importantly, these pathways contain only bridging relationship information (structure), and not the bridging concepts themselves (content).
and 234 negative test cases. Each test case is a drug-ADE pair, and each drug is one of 181 unique drugs in the set, including NSAIDs, beta-blockers, ACE inhibitors, antidepressants, antibiotics, and more. The four side-effects chosen – acute myocardial infarction, acute renal failure, acute liver failure, and gastrointestinal bleeding – are four of the most significant ADEs for a risk identification system. Together, these combinations provide a widely used methodological evaluation set and the current benchmark in PV.

Methods

To facilitate our research, we utilized the semantic predications extracted from the literature by SemRep housed in the Semantic Medline Database (SemMedDB) version 2.2 database and the 2012 MetaMapped Medline Baseline (MBB) for PSI and co-occurrence approaches respectively. The SemMedDB extractions were generated by SemRep version 1.5. The MMB was derived from 20,494,848 citations up to November 2011, and contains 399,701 distinct concepts, while SemMedDBv2.2 was derived from 22,252,812 citations up to March 2013, and contains 63,795,467 predications spanning 58 predicates and 257,350 distinct concepts. These versions are identical to those used in the previously published analyses, and were chosen to facilitate methodological comparison. As in previous work, negative predications, such as drug DOES NOT TREAT side-effect, were excluded, comprising only 1.2% of the total predications.

Of the OMOP data set, we utilized 164 positive and 230 negative test cases. For our analysis, we did not use test cases for the drugs darunavir and sitagliptin, as they did not occur in the vector representation stores used for the analyses. The four side-effects in the OMOP set can be defined by a list of International Classification of Diseases (ICD) 9 codes, and so we expanded the list of terms encompassing each ADE to all of its ICD-9 codes and sub-codes. Table 2 represents some of the expanded query terms used for myocardial infarction. Drug names were not expanded, and were queried in all cases as named in the OMOP reference set except for niacin, which was translated to nicotinic acid.

<table>
<thead>
<tr>
<th>OMOP Term</th>
<th>ICD-9 Code</th>
<th>Expansion Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>acute myocardial infarction</td>
<td>410</td>
<td>acute_myocardial_infarction</td>
</tr>
<tr>
<td>411</td>
<td></td>
<td>acute_coronary_syndrome</td>
</tr>
<tr>
<td>414</td>
<td></td>
<td>silent_myocardial_infarction</td>
</tr>
</tbody>
</table>

Table 2. Example expanded terms for myocardial infarction.

As our co-occurrence approach, we employed reflective random indexing (RRI) which considers both direct co-occurrence and indirect relatedness (between terms co-occurring with the same other terms). RRI was implemented using the Semantic Vectors package version 3.7 with 32,000 dimensional binary vectors in accordance with the BSC. Briefly, document vectors are built by superposing the elemental vectors for each distinct concept that occurs in each document in the MMB, using a log entropy weighting metric (which reduces the effect of high-frequency terms across and within documents). Semantic concept vectors – i.e. S(concept) – are then built by superposing the document vectors that the given concept occurs in. These semantic vectors are then rank ordered based on (1 – the normalized Hamming distance), a similarity measure, between drugs and (expanded) ADEs in the reference set. To expand each ADE into a query that reflected the sum of its ICD-9 codes, we superposed the available vector representations of expanded list terms reflecting the given condition.

For both producing results as in previous PSI analyses using discrete pathways (i.e. DbA) and for our classification analysis, we utilized vector stores from previous analyses, whereby 32,000 dimensional binary vectors were generated consistent with the BSC using Semantic Vectors version 3.711. A maximum frequency threshold of 1M terms was used to prune uninformative high-level concepts, and negative predications were removed as mentioned above. In order to further mitigate the effect of highly frequent, uninformative terms, superposition of bound products were weighted by predicate frequency multiplied by the sum of the inverse document frequency of the predicate and bound concept such that (using ibuprofen as an example):

\[ S(\text{ibuprofen}) \leftarrow E(\text{TREATS}) \otimes E(\text{pain}) \cdot f_{\text{TREATS}} \cdot (idf_{\text{TREATS}}+idf_{\text{pain}}) \]

where \( f_{\text{TREATS}} = \log(1+\text{occurrences of prediction ibuprofen-TREATS-pain}) \)

\( idf_{\text{TREATS}} = \log(\text{number of total predications} / \text{number of predications containing TREATS}) \)

\( idf_{\text{pain}} = \log(\text{number of total predications} / \text{number of predications containing pain}) \)
### Table 3.
The double predicate discrete reasoning pathways that make up the subspace for DbA. gih = gastrointestinal hemorrhage.

<table>
<thead>
<tr>
<th>Predicate Subspace Component</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(INTERACTS_WITH) ⊗ P(CAUSES-INV)</td>
<td>Valdecoxib interacts with something that causes gih.</td>
</tr>
<tr>
<td>P(ASSOCIATED_WITH) ⊗ P(COEXISTS_WITH)</td>
<td>Valdecoxib associated with something that coexists with gih.</td>
</tr>
<tr>
<td>P(COMPAred_WITH) ⊗ P(CAUSES-INV)</td>
<td>Valdecoxib compared with something that causes gih.</td>
</tr>
<tr>
<td>P(ASSOCIATED_WITH) ⊗ P(INTERACTS_WITH)</td>
<td>Valdecoxib associated with something that interacts with gih.</td>
</tr>
<tr>
<td>P(ISA) ⊗ P(CAUSES-INV)</td>
<td>Valdecoxib is a type of something that causes gih.</td>
</tr>
</tbody>
</table>

Additionally, ADE terms were expanded as in RRI for DbA and CbA. For each drug/ADE pair in the reference set, the semantic vectors for the respective drug and ADE are then released as shown above (Table 1), giving us the abstract pathway vector representation for that pair. This vector is then projected into a subspace composed of two predicate discrete reasoning pathways (Table 3) and rank ordered by magnitude to generate DbA results, and passed as input to the machine learning algorithms in our CbA work. Only double predicate pathways were utilized for DbA to facilitate a fair comparison with CbA (in previous work, triple-predicate pathways were also considered).

With our goal to classify these vectors according to ground truth relationship, we labeled the vectors representing each drug/ADE pair relationship with the OMOP ground truth state (Figure 1) for input into machine learning algorithms.

---

All machine learning development and deployment was done in the Python programming language version 3.5 using the scikit-learn package. The specific Python deployment was from Continuum Analytics Anaconda platform, version 2.5, and a development environment was created using the conda command line utility packaged with Anaconda specifically for the purpose of these analyses and to ease reproducibility. This environment file, along with saved data arrays, code files used, and additional information on software versions utilized, are available upon request.

Within scikit-learn, we utilized the LibLinear library through the scikit-learn.LinearSVC() front end; for k-Nearest Neighbors (kNN) we utilized the scikit-learn.kNN() front end; and for LR we utilized the LibLinear library through the scikit-learn.LogisticRegression() front end. Hyperparameters for the SVM (the regularization C parameter), for LR (C parameter) and for kNN (the k number of neighbors) were chosen using cross-validation grid search functionality built into the scikit learn package (scikit-learn.GridSearchCV). For the SVM and LR, an L1 penalty parameter was chosen to enforce sparsity in the learned model. ROC AUC curves for SVM and LR models were generated by passing the results of the decision function, which calculates the distance from the dividing hyperplane.
of the classifier for each example in the case of the SVM, as the rank ordering. Additionally, Stratified 5-Fold cross-
validation was used to generate the mean ROC AUC curve by averaging each fold’s performance on a held out test
set. Learning curves were generated for the F1 metric by varying the number of training examples and the
dimensions of the data. Test data was never utilized in the training phase of any supervised machine learning
approach in an attempt to mitigate over-fitting. All plotting was done using the matplotlib package in Python.

Results

A comparison of kNN performance results with 5, 10, and 15 nearest neighbors are reported below in Table 4. kNN
typically performed best with $k=5$ in our analysis. For comparison, see F1 scores in Figure 4.

<table>
<thead>
<tr>
<th>k Neighbors</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.88 +/- 0.06</td>
<td>0.87 +/- 0.06</td>
<td>0.87 +/- 0.06</td>
</tr>
<tr>
<td>10</td>
<td>0.86 +/- 0.05</td>
<td>0.86 +/- 0.05</td>
<td>0.86 +/- 0.05</td>
</tr>
<tr>
<td>15</td>
<td>0.83 +/- 0.05</td>
<td>0.81 +/- 0.06</td>
<td>0.82 +/- 0.05</td>
</tr>
</tbody>
</table>

Table 4. kNN precision, recall, and F1 scores reported with two times the standard deviation across cross-
validation runs in error per number of k neighbors used.

A summary of the ROC results between DbA, RRI, and SVM and LR models both utilizing a C of 1, including
AUCs, can be found in Figure 2. All variants of CbA outperformed DbA (+≈37% over 0.68 AUC) and RRI (+≈48%
over 0.63 AUC) models, with AUCs around 0.93-0.94. Additional context for these results can be found in the
learning curves presented in Figure 4.

For C=1, SVM coefficient vectors contained approximately 300/32000 nonzero weights when trained, with similar
levels of nonzero weights in LR models. Coherence of nonzero features, including sign and strength of weight, is
shown in Figure 3. These nonzero features are distributed across the coefficient vectors for these models. There are
approximately 70% shared nonzero positions between them.

Figure 2. ROC Curves for tested models. AUCs are listed +/- two times the standard deviation (across cross-validation runs).
LR and SVM models both have one hundred percent accuracy on training sets when trained with full dimensional vectors with $C=1$. Cross-validation performance is consistently above 0.80 when considering an F1 scoring metric with even a small portion of the training data set and significantly diminished input vector dimensions.

**Figure 3.** Coherence among LR and SVM models. Negative weights are colored in white, zero values colored in grey, and positive values colored in black. In either case, the 32k weights - one for each dimension of input data - are reshaped into a 200x160 matrix for visualization. Nonzero elements are distributed across the weight vectors. Contrast adjusted for visibility.

LR and SVM models both have one hundred percent accuracy on training sets when trained with full dimensional vectors with $C=1$. Cross-validation performance is consistently above 0.80 when considering an F1 scoring metric with even a small portion of the training data set and significantly diminished input vector dimensions.

**Figure 4.** The effect that truncating dimensions (a) and limiting training samples (b) has on cross-validation performance of CbA parametric methods, as measured by an F1 metric.

**Discussion**

To our knowledge, these results represent the first evaluation of classification based on vector representations of abstract relational structures. kNN performance indicates that relational neighborhood as defined by abstract predicate pathways is of value for this task. Although we believe that performance metrics indicate the LR and SVM models are over-fitting, the strong cross-validation performance implies remarkable self-consistency in the representations of the literature for the OMOP reference set. Additionally, LR and SVM models both have striking
coefficient vector homology. Since the OMOP set derived its positive examples using automated NLP methods to extract drug reactions from package inserts as the first pass, and since negative examples were likewise selected as those that did not show up in drug packaging inserts, Tisdale evidence, and had nonexistent or only negative linkings to the ADE in question in the literature during manual review, it is possible that their selection criterion generated a reference set that is internally coherent amongst examples of each class. While more work needs to be done to assess generalizability, presented results indicate that there exists a self-consistent structure to the OMOP reference set as represented here that provides CbA approaches sufficient information to make accurate classifications on it.

The machine learning algorithms we chose are particularly simple and incorporate no prior information; we anticipate the incorporation of priors will improve performance and likely lend to more generalizable models. Unsurprisingly, it also seems that the models would be well served with additional examples, as evidenced by the learning curves. Such examples would ideally introduce more variance, and trained models would be less prone to over-fitting. Experimentation with additional weighting strategies within the representation itself may also be warranted, as the weighting procedures can influence which predicate pathways are encoded more influentially.

In addition to only being characterized in the context of a single data set, our study has other limitations worth mentioning. First, we didn’t attempt to optimize any statistical weighting metrics or other PSI parameters in the development of this work. Second, we didn’t systematically evaluate how performance changes cross different initializations of the random vectors, opting instead for pseudo-random instantiation per [33]. Additionally, work is presented in comparison to similar approaches that operate on literature information exclusively; other studies have been published on the data set utilizing different methods and data, including observational data34,35, which is not considered here.

One major benefit of the DbA approach, and one current limitation to CbAs, is clearly defined interpretability. For DbA, confidence scores are generated based on vector-subspace similarity, and the subspace is constructed from elected double predicate pathways. If something has high similarity to the subspace, then the interpretation is that it has high similarity to those predicate pathways which make up the subspace. In our CbA approaches, it is difficult to directly map nonzero coefficients to interpretable pathways or features. Since a fully distributed representation encodes information across vectors, individual features are primarily important in their context of other features. The primary challenge we see facing this work is in mapping learned nonzero parameters from these classifiers back to more interpretable predicate pathways, and to the literature sources that plausibly link these entities themselves. In this same vein, it is likely that more information than just common relations are being utilized by classification algorithms as presented here. The vector space, as a function of its structured encoding scheme, likely incorporates common object information as well when comparing our derived relational representations. For example, in comparing drugs A and B that inhibit the same target C, A\(\bowtie\)C would be similar to B\(\bowtie\)C, and is functionally similar to a direct object comparison in that context (i.e. similarity(A\(\bowtie\)C, B\(\bowtie\)C) = similarity(A,B)). The information used in this case is not exclusively relational, though relational information plays a role (A inhibits y, A causes C; B inhibits Y therefore B may cause C). Further research is needed to identify the extent to which these different sorts of information contribute toward classifier performance.

Conclusion

Each model accurately predicted across a variety of drugs and four ADEs, learning only one set of parameters to distinguish between plausibly causal as defined by the OMOP set and not, substantiating that the representation itself is a meaningful basis for classification tasks. Our original hypothesis that the basis of the relational structure between pairs of biomedical entities could provide the necessary information for categorization of higher level relational status was substantiated by our results on the OMOP reference set.

Acknowledgements

This work is supported by a training fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (Grant No. T15LM007093) and U.S. National Library of Medicine Grant (1R01LM011563), Using Biomedical Knowledge to Identify Plausible Signals for Pharmacovigilance.
References

12. Hristovski D, Burgun-Parenthoine A, Avillach P, Rindflesch T. Towards using literature-based discovery to explain drug adverse effects. 24th International Conference of the European Federation for Medical Informatics Quality of Life through Quality of Information. MIE; 2012.
Healthcare Data Analytics for Parkinson’s Disease Patients: A Study of Hospital Cost and Utilization in the United States

Sunanda Mukherjee, MS\textsuperscript{1}, Huanmei Wu, Phd\textsuperscript{1,2}, Josette Jones, Phd\textsuperscript{1}
\textsuperscript{1}School of Informatics and Computing, IUPUI, Indianapolis, IN, USA;  
\textsuperscript{2}Department of Computer and Information Technology, IUPUI, Indianapolis, IN, USA; 

Abstract

Parkinson’s Disease (PD), a prevalent problem, especially for the aged populations, is a progressive but non-fatal nervous system disorder. PD patients have special motor as well as non-motor symptoms over time. There are several limitations in the study of PD such as unavailability of data, proper diagnosis and treatment methods. These limitations significantly reduce the quality of PD patient life quality, either directly or indirectly. PD also imposes great financial burdens to PD patients and their family. This project aims to analyze the most common reasons for PD patient hospitalization, review complications that occur during inpatient stays, and measure the costs associated with PD patient characteristics. Using the HCUP NIS data, comprehensive data analysis has been performed. The results are customized visualized using Tableau and other software systems. The preliminary findings sheds light into how to improve the life quality of PD patients.

Keywords

Parkinson’s disease, Data Analytics, Hospitalization utilization and cost, Tableau, HCUP NIS database

1. Introduction

Parkinson’s disease (PD) is a slowly progressive, chronic but non-fatal neurological movement disorder or disorder of the brain. It is the second most common neurodegenerative disease in US. It is reported that there is an average of 0.3\% of PD prevalence throughout the entire population as estimated by the individual studies based on the projection of economic burden of PD\textsuperscript{1}. Previous study also showed there was about approximately 2\% and 5\% increase rate in the number of patients identified with PD over age 65 and over 85, respectively. In addition, the overall prevalence and global economic burden of PD in World’s top 10 densely populated nations has been investigated and it is estimated that the PD occurrence rate will be double by 2030.\textsuperscript{2}

Although the actual cause of PD remains unclear, it has been seen that the symptoms of PD is associated with the deficiency of Dopamine, a chemical in the brain. There is no explicit tests to diagnose PD. Thus, the disease is diagnosed only based on the patient’s report and the combination of exhibited symptoms. Confirmed PD is typically treated with the combinations of medications and exercises. Medicines can improve the symptoms to some extent, but the care and management is very challenging and costly for the patients and their healthcare providers.

The major symptoms of PD disease is characterized by a collection of motor as well as non-motor symptoms, which is the results from dopaminergic neuron degeneration.\textsuperscript{1,3} The major PD motor symptoms is clustered into an acronym called TRAP: Tremor, Rigidity, Akinesia or Bradykinesia and Postural instability. The non-motor symptoms include mood disorders, sleep disorders, depression, anxiety and decreased motivation etc. These symptoms have a huge influence to the potential adverse effects on health-related life quality of PD patients. The severity of such symptoms indicates the progression of the disease, which resulting in a higher costs of medical treatments for hospitalization and care coordination.

The major objectives of this study are to perform data analytics, to highlight the most common reasons for PD patient hospitalization, to determine hospitalization costs, and to measure the co-morbidities associated with this disease. The long term goal of this study is to lessen the symptoms of PD and hence, improve the quality of life of the patients by lowering the hospitalization costs and inpatient hospital stays.

2. Materials and Methods

2.1 Data Source:

A retrospective analysis was performed using the Healthcare Cost and Utilization Project's (HCUP) Nationwide Inpatient Sample (NIS)\textsuperscript{4} data for the year 2012. The Agency for Healthcare Research and Quality (AHRQ) sponsors
the HCUP databases that are specifically designed to determine and identify patterns in hospital utilization and cost across the United States. The HCUP-NIS database is the largest inpatient database available in the US, which represents a sample of non-federal US community hospitals. Just in 2012, there are 1,049 hospitals from 46 states in US participated in NIS projects. More than 7 million records of hospital stays were reported by these hospitals. It is estimated that the sampling includes over 95 percent of discharges from the hospitals participated in NIS. The estimated samples were weighted in order to minimize the margin of error and to reflect all 50 states across the US. The large sample size of the database enabled us to analyze rare conditions, uncommon treatments and special patient populations.

To protect the privacy of individual patients, physicians, and hospitals, the state and hospital identifiers are de-identified. There are many clinical and non-clinical hospitalization data elements recorded in the HCUP NIS database. Sample non-clinical information are patient demographic information, payment related information, hospital characteristics (such as region, census division, and location), and total charges. Sample clinical related information include principal and secondary diagnosis and procedure (includes both ICD-9 and CCS codes), discharge status and Length of stay.

2.2. Variables of Interest:

At first, based on the ICD-9-CM diagnosis codes, the individuals who were primarily diagnosed as PD patients at the time of admission had been identified. In HCUP databases, more than 14,000 ICD-9-CM diagnosis codes and 3,900 procedure codes had been mentioned which were further classified and clustered into lesser number of clinically appropriate categories by the AHRQ's Clinical Classification Software (CCS). This enabled the database to capture a large population of relatively similar conditions into a single group by making the information more useful for performing statistical analyses and developing reports. PD patients were identified using CCS diagnosis code as 79.

One of the dependent or outcome variable of interest was the total hospitalization cost resulting from the entire hospital stay. The total hospital charges counted all the expenses for services provided to the patients excluding physician or professional fees.

For the analysis, among the predictor variables, the primary and secondary disease conditions and procedures and length of the hospital stays were important. The preliminary diagnoses had been identified at the time of admission whereas the comorbidities or the secondary disease conditions were recorded throughout the entire hospital stay. The similar manner has been followed for the determination of the primary and secondary procedures. The length of stay was the time period from the admission to the discharge of latest hospital stay of each patient.

2.3. Approaches:

Retrospective analysis was performed over the HCUP-NIS database focusing on the determination of the hospital utilization cost for PD patients. Descriptive statistics were used to summarize the results. The mean and standard deviation was used to explain the continuous variables. On the other hand, the categorical variables were shown in percentage values.

Statistical approaches have been used to determine the correlation of clinical features including diagnosis, current medical conditions and costs. For the Univariate analysis, the two major characteristics of the variable analyzed were the distribution and the central tendency.

Data were analyzed using R language and Tableau software. The Tableau software provides a great user interface and visualization tools for the analytical results.

3. Results & Findings

Extensive analysis were performed for the 2439 PD patient hospitalization records. The preliminary results are presented below.

3.1 Demographics:

Age Distribution: The patients were distributed according to their age group as shown in the (Figure 1). The result shows that a total of more than 99% of patients were in the range of 50 to 90 age group where approximately 26%, 34% and 25% patients were in 61-70, 71-80, and 81-90 age groups, respectively. Although, in earlier studies, there is no significant association has been found between the aging process and the acceleration of the disease, still it is a debatable fact whether the aging influences the progression of the disease or not.
Figure 1. The distribution of PD patients by the age group.

Patients across the census division: The data is geographically divided into 9 divisions from 4 regions. Figure 2 below shows all the states in each of the 4 US Census Bureau Regions and 9 divisions. The states highlighted in grey color do not participate in HCUP.

Figure 2. The map depicting the 9 census divisions and 4 regions in US.
Figure 3. The percentage of PD patients across the census division and the region in US.

For the four geographical divisions, the Middle Atlantic division in the Northeast Region contributes to the highest number of PD patient sample (20% of total 2439 patients) whereas the New England division from the same region contributes to the lowest number of PD patient sample (5% of total 2439 patients) as described by the (Figure 2). Next to the Middle Atlantic division, the South Atlantic division from South region, East North Central division from Midwest region and Pacific division from the West region contributes to the more number of inpatient samples.

3.2. Diagnosis Related Group (DRGs):

DRG is a system of classification that has been developed with the intent to identify the presence of complications or comorbidities and to replace "cost based" reimbursement that had been used up to that point by the hospitals. In contrast, according to the literature by Muller et al.\textsuperscript{12} published in 2010, a significant decrease in the reimbursement of treatment costs has been noticed in spite of the presence of increased operation risk and medical conditions that in turn results in an increased economic burden for the PD patients.

The DRGs, identified for the PD patients, included craniotomy and endovascular intracranial procedures, spinal procedures, extracranial procedures, peripheral/ cranial nervous system procedures and degenerative nervous system disorders. The DRGs also indicated the severity of the disease as it has identified the patients with or without minor or major complications and co-morbidities.

The top 6 DRGs having the percentage of the total number of PD patients in each group is shown in (Figure 4). The description of these DRGs are -

- Degenerative Nervous System Disorders W/O MCC (57),
- Craniotomy & Endovascular Intracranial Procedures W/O CC/MCC (27),
- Degenerative Nervous System Disorders W MCC (56),
- Cranio W Major Dev Impl/Acute Complex CNS PDX W/O MCC (24),
- Craniotomy & Endovascular Intracranial Procedures W CC (26)
- Periph/Cranial Nerve & Other Nerv Syst Proc W/O CC/MCC (42)

where W CC, W MCC, W/O MCC and W/O CC/MCC corresponds to With Complications or Co-morbidities, With Major Complications or Co-morbidities, Without Major Complications or Co-morbidities, Without Complications or Co-morbidities/ Major Complications or Co-morbidities respectively.
3.3. Frequently Used Procedures:

The study has identified total top 11 primary procedures that has been frequently used for the PD patients. The Figure 5 shows 10 primary procedures having percentage frequency in between 0.5% to 2% and excludes the procedure with ICD-9 code as 293 having the highest frequency of 20%. The Table 1 listed the top 10 primary procedures based on
their average cost per day. Considering both the Figure 5 and Table 1, the study found it interesting that the procedures having higher average cost per day has been frequently used for the PD patients. This analysis has also listed 3 interesting procedures based on their cost and usage frequency:

1) The procedure Imp/repl brain stim lead (ICD code 293): It having the highest frequency, is used for 20% of the total patients. The description of the procedure is known as Implantation or replacement of intracranial neurostimulator lead(s). This is applied to the implantation, insertion, placement, or replacement of intracranial brain pacemaker [neuropacemaker], depth electrodes, foramen ovale electrodes and intracranial electrostimulator.

2) Implantation or replacement of peripheral neurostimulator lead(s) (ICD code 492): This procedure having the highest average cost per day is also used frequently for the patients.

3) The cranial implantation or replacement of neurostimulator pulse generator (ICD code 120): It is a procedure which is also found in the top 10 list, and is followed along with the above procedure (ICD code 293).

Table 1. The top 10 procedure codes, description of procedure, average cost/day and standard deviation

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Description</th>
<th>Average cost/day</th>
<th>Standard Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>492</td>
<td>Imp/repl peri stim lead</td>
<td>64636.29</td>
<td>30449.76</td>
</tr>
<tr>
<td>8695</td>
<td>Ins/re pls gn no rechrg</td>
<td>57862.33</td>
<td>23096.03</td>
</tr>
<tr>
<td>293</td>
<td>Imp/repl brain stim lead</td>
<td>50798.37</td>
<td>27336.41</td>
</tr>
<tr>
<td>120</td>
<td>Imp/repl brain pulse gen</td>
<td>39664.58</td>
<td>2474.76</td>
</tr>
<tr>
<td>8841</td>
<td>Contr cerebr arteriogram</td>
<td>19949.33</td>
<td>17334.49</td>
</tr>
<tr>
<td>8659</td>
<td>Skin closure NEC</td>
<td>13138.78</td>
<td>10156.98</td>
</tr>
<tr>
<td>8919</td>
<td>Video/radio eeg monitor</td>
<td>9873.53</td>
<td>4955.67</td>
</tr>
<tr>
<td>9339</td>
<td>Physical therapy NEC</td>
<td>9572.88</td>
<td>660.61</td>
</tr>
<tr>
<td>8944</td>
<td>Cardiac stress test NEC</td>
<td>9269.11</td>
<td>3877.17</td>
</tr>
<tr>
<td>4525</td>
<td>Clos large bowel biopsy</td>
<td>8925.88</td>
<td>7322.97</td>
</tr>
</tbody>
</table>

Figure 6. The length of hospital stays for PD patients
3.4 Length of Patient Stay:

The average length of stay for the top 10 primary procedures was approximately 4.8 days, with a range of 1-16 days. Qualitatively, the PD patients treated with the procedure called percutaneous endoscopic gastronomy had the longest stays whereas the patients treated with insertion of catheter had the shortest stays as described by the Figure 6.

3.5. Average cost per procedure:

The study calculated the average cost for each of the primary procedures. The below diagram shows the plots for top 15 primary procedures and each boxplot represented one primary procedure as shown in Figure 7. The plots were sorted based on the mean value of cost per day in an ascending order.

The average cost for all the top 10 procedure was approximately $28400 (range ~ 9000-65000) with the standard deviation as approximately $12800 (range ~ 660-30450). The below table (Table 1) shows the average cost and standard deviation for the top 10 primary procedures.

![Figure 7. The average hospitalization cost for top 15 frequent procedures](image-url)
When the study compared the procedures by taking average cost, average length of stay and frequency into consideration, it has identified the below procedures for which the average cost, length of stay and the frequency was higher as described by the (Figure 8) -

- Imp/repl brain pulse gen (120)
- Imp/repl brain stim lead (293)
- Imp/repl peri stim lead (492)
- Percu endosc gastrostomy (4311)

### 4. Discussion:

This analysis of population-based hospital discharge data from patients with PD reveals variations in hospital costs by patient demographic characteristics. Average hospital cost was significantly higher for hospitalizations. Several important observations were generated. A major component of the overall economic burden of healthcare is the initial cost. The study identified the procedures, having higher cost and length of stay, was related to the Deep Brain Stimulation (DBS) which is an important treatment option for PD patients experiencing dyskinesia and motor complications. Literature related to DBS showed that the most common reasons for prolonged hospitalization after DBS are mental status changes and cerebral hemorrhage.\(^{13,14}\)

Although the study illustrated the large economic burden of PD in terms of medical costs, still it couldn’t able to demonstrate the true burden of the disease. The presence of comorbidities and its treatment procedures had been significantly reduced the quality of life for people with PD and their family members. It has been further projected that the pressure of such chronic conditions will substantially increase in next few years with the increase of size of the aged populations.\(^{15,16}\) Such projections demonstrate the need for the innovation of old techniques and identification of new treatment procedures including vigorous exercises in order to prevent or delay the onset of the disease; hence, increasing the quality of life for the PD patients.

The present study had several limitations common to administrative databases. It was susceptible to errors for categorizing patients and calculating costs. The hospital charges analysis does not constitute the hospital-specific
differences in charge practices. There may be variation in cost-to-charge ratio (CCR) by patient population and by individual hospital departments that could have resulted in errors in calculating costs. Thus, it is difficult to develop comprehensive framework of current payment models focusing on cost structure on PD patients because up-to-the minute charge data are not available, and charge practices may be lost in the heterogeneity of hospital type.

However, based on these data, a predictive model of cost can be built in future so that the model can be additionally utilized in the cost management debate and the creation of data-driven policies. This can fuel further studies in the field and provide elements for the design of prospective investigations.

5. Conclusion

This study has successfully identified the most common reasons for hospitalization in PD, determined costs associated with nationwide hospitalizations of individuals with PD and measured the co-morbidities associated with the disease using the nationwide discharge sample data across the US. Future work, incorporating the development of a prediction model of hospitalization utilization cost can fuel the further studies in the field of healthcare.

References

Medication Harmony: A Framework to Save Time, Improve Accuracy and Increase Patient Activation

Frank Pandolfe MD,1,2 Bradley H Crotty MD MPH,1,2 and Charles Safran MD MS1,2

1Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA
2Harvard Medical School, Boston, MA

Abstract

Incompletely reconciled medication lists contribute to prescribing errors and adverse drug events. Providers expend time and effort at every point of patient contact attempting to curate a best possible medication list, and yet often the list is incomplete or inaccurate. We propose a framework that builds upon the existing infrastructure of a health information exchange (HIE), centralizes data and encourages patient activation. The solution is a constantly accessible, singular, patient-adjudicated medication list that incorporates useful information and features into the list itself. We aim to decrease medication errors across transitions of care, increase awareness of potential drug-drug interactions, improve patient knowledge and self-efficacy regarding medications, decrease polypharmacy, improve prescribing safety and ultimately decrease cost to the health-care system.

Introduction

Patients receive prescriptions from multiple clinicians and often use more than one pharmacy. They may take over-the-counter (OTC) medications and herbal supplements. For patients and clinicians alike, constant reconciliation of a medication list has been a Sisyphean task. Discrepancies and inaccuracies among medication lists are linked to errors in prescribing and self-administration. Clinicians often settle for a list that is “good enough,” yet contains numerous errors. These errors translate into adverse drug events (ADEs) and often result in increased morbidity, mortality, hospital readmission, and increased health-care-related cost1-4. In the United States, the cost of ADEs is estimated at $76.6 billion in the ambulatory setting alone5.

Numerous studies have assessed how to decrease ADEs, and hospital readmission rates6-11. Some form of intervention to obtain the “best possible medication history” over typical care has repeatedly shown to decrease error rates on the medication list. Research to date has primarily focused on improving the accuracy of a singular medication list and decreasing ADEs. Thus far, however, there has not been much consideration to the amount of institutional infrastructure and person-time needed to complete such tasks and how often the task must be repeated.

Some studies have suggested cost-benefit by paying pharmacists to reconcile the best possible medication history upon hospital admission, a process that can take more than 30 minutes per patient12. Broader system changes include creating leadership positions, changes in workflow, management teams, “reconciliation champions”, interdisciplinary teams, encouragement from those who are “less than enthusiastic” to participate, and other daunting barriers to implementation13,14.

Currently, the “gold standard” of obtaining the best possible medication history involves a pharmacist compiling patient history, obtaining patient and office medication lists, pharmacy data, and discharge summaries upon admission. Approaching discharge, the pharmacist reviews changes in medications, describes indications and reviews discontinued medications15,16. Post discharge, the pharmacist contacts the patient again to resolve any unanswered questions. Many institutions are not yet current with these best practices or have limited pharmacy staff to complete such tasks. The process of medication reconciliation is then delegated to nurses or physicians who often do not have time or resources to ensure a high-degree of accuracy. Although this workflow is effective for a single hospitalization, it is a laborious task that must be repeated at every transition of care and does not address the issue of creating multiple out-of-date medication lists.

Medication managers through mobile applications and institution-based systems currently exist, however, under fractured infrastructure17-19. A hurdle for federation is the siloing of information in institution-specific systems (Figure 1). These systems fail to communicate with one another and patients are often responsible for the input of all
medications by hand. Even with interoperability and communication among systems, keeping multiple lists synchronized remains challenging.

**Figure 1.** Current structure of the outpatient medication list

Although an approach such as “Blue Button” allows patients read-only access to their own information, a user must be actively engaged and obtain separate log-in information and credentials at each institution they receive care. Additionally, patients do not have the ability to easily modify the data provided by “Blue Button”. A consolidated, shared, patient- and physician-modifiable view of a medication list hasn’t been successfully implemented.

To achieve provider buy-in we need to overcome the tremendous time and effort needed to keep an accurate medication list. We must increase patient activation, and where appropriate, family members, by offering them utility in keeping a list. For most patients, the list represents an array of unpronounceable names of various colored pills that he blindly takes. We must incorporate common-language, actionable information while minimizing patient effort for information loading and updating of their medications.

**Envisioning a Solution**

A solution should allow patients, providers, institutions and pharmacies to eliminate disparities among a medication list that is used across various environments. It must be:

- Widely visible while protecting patient privacy
- Modifiable by patients, caregivers, and practitioners, and family members where appropriate
- Institution agnostic
- Able to easily fit into the current framework of medication reconciliation so that minimal disruption in workflow is caused.

The solution must also encourage patient activation and aid in comprehension. Therefore, it should:

- Easily capture OTC medications
- Pre-populate the list to avoid confusion and fatigue
- Provide knowledge regarding drug information, including pill images
- Alert for potential interactions

A benefit to such a solution is a faster and more accurate medication reconciliation process thus reducing ADEs from inadvertent prescribing errors.
Using the HIE to centralize the medication list and establish a “single source of truth”

Practically speaking, a dedicated, stand-alone database and accompanying application to act as a medication manager must be implemented. For many locations, a HIE is established to help share data among various providers. As a hospital or a physician office registers its EHR as a provider on a HIE, so too will the medication manager. By doing so, it will be able to interact with other “trusted” systems on the network. Its job will be to maintain a database of an individual’s medications by gathering information from other institutions that have recorded medications (Figure 2). Linking the medication manager into the HIE enables a dedicated provider of medication information. Furthermore, providers such as PCPs, specialists, hospitals and nursing homes will have access and the ability to update the list.

Under the proposed solution, a patient would register for this service via a web site or a smart-device application. Identity would be verified and the user authenticated. Exchange of health information over a HIE requires patients to accept the terms and conditions of their local HIE. For example, Massachusetts has an opt-in policy for patients. Consent is possible online and one can opt-out at any time. This process can be accomplished within the medication manager app itself.

![Figure 2. Centralized Medication Manager](image)

**Interoperability**

To be viable, the solution must be able to fit within current workflow situations. Although difficult to imagine within our current EHR landscape, effort is currently underway towards interoperability. We must be forward thinking and design for such a future. By using the same infrastructure an EHR uses to pull information from a Surescripts® Medication History data source, the EHR would also be able to pull information from the medication manager. In addition, as an EHR is able to send an e-prescription to a pharmacy, it would also be able to send a message to the medication manager. The medication manager would have the capability to consume HL7-NCPDP SCRIPT e-prescriptions, the National Council for Prescription Drug Programs standard for transmission of...
prescription information\textsuperscript{22-27} to update information. In the future, it could also run as a SMART application\textsuperscript{28} utilized as a web-app, mobile device-app or trusted service within an EHR.

**Data federation and reconciliation**

Once verified, medication information would be federated from available sources (Figure 3). Pre-population of data can be accomplished by gathering data from inpatient and outpatient visits via EHR generated Consolidated-Clinical Document Architecture (C-CDA) documents. Pharmacy data can be gathered from Surescripts data. It is possible that at some point in the future, pharmacies would be willing to expose patient fill information onto the HIE as well, providing an additional data source.

Once the data is collected, a matching algorithm would be applied to find overlaps and eliminate duplicates\textsuperscript{29}. Patients would have the ability to verify and change any of the listed medications. In addition, the patient would be able to add other prescribed medications not appearing on any lists (i.e. from providers not participating in the HIE or older, yet still active, medications). Over-the-counter medications, less commonly captured by traditional methods, can be added via bar-code scan, text-matching input or optical character recognition (OCR) software.

Updating the medication list can occur at transitions of care, patient-provider interactions, or at any time by the patient. Primary users of this system will be providers who will rely on it to maintain an accurate list of medications that can be shared and updated by other providers; health-care institutions, who will use it to more easily reconcile medications during points of transition; and patients who will be able to add, delete, modify and verify OTC and prescribed medications. Each time a patient or provider logs in, a query will be performed and if information more recent than the last time of reconciliation is discovered, it will be presented for reconciliation. This way, every time the list is accessed, a pro-active and pre-population approach to reconciliation is taken. Doing so, will increase accuracy and fidelity of the medication list in addition to preventing fatigue from demanding that the list be created from scratch. The combination of these services will make it easier for patients to have a list that reflects the true medications that they are taking.

*Figure 3. List federation and reconciliation*
Encouraging patient engagement; using an interactive interface that provides information

Patients are left to their own devices to maintain a medication list. Either they write out a list (which is often out of date and lacks vital information such as dose and frequency), use a list printed from their PCP office (often containing cross-outs, corrections, omissions and out of date information), bring a bag of pills, or rely on their memory. Such solutions typically don’t provide access to other useful information such as images of medications, indications, drug class, or potential interactions. Apps that focus on adherence typically require that the patient hand enter all medications. These applications are vendor specific, and although it may be able to print or show a list to others, there is not a standardized way to integrate that information across platforms. For the elderly or medically complex patient who may be on multiple prescription and over-the-counter medications, the full medication list may not be captured sheerly due to the initial time investment needed to complete the process.

Using government resources, such as RxNorm, MedlinePlus, and RxImage provided by the National Library of Medicine, a well-designed medication manager can provide services such as brand and generic name, dose, frequency, images of the medications, common indications, convenient links for more detailed drug information, and an interaction checker. The combination of these services will aid in patient-education in addition to recognizing possible unintentional polypharmacy. In addition, the medication manager will provide an opportunity for dialogue regarding differences in the way that medications are prescribed and taken. Discontinued medications can be moved to a separate list keeping track of prior medications (Figure 4).

![Figure 4. Patient view of current medications](image)

**Time and safety benefits**

When a patient presents to the hospital or doctor’s office, providers attempt to fulfill the Joint Commission’s requirement to “maintain and communicate accurate patient medication information”, and thus undertake the process of formulating the best possible medication history. As mentioned, this is time-consuming, and non-scalable job that includes collecting lists from various sources. In addition, errors are common and often result in ADEs. Under the proposed solution, these lists would be instantly and electronically obtained and compared. (Figure 3) The job of formulating an accurate medication list is simplified to resolving discrepancies among the lists.
On admission, a decision regarding whether to stop, continue or modify each medication is made. The proposed system would not be involved with medication management during an inpatient stay. On discharge, the admission list is compared to the inpatient list and the discharge medication list is formulated. As described above, the discharge medication list would be available to the centralized medication manager either via an e-prescription or C-CDA document and accessible by outpatient providers and the patient, decreasing confusion regarding an accurate discharge medication list. Available drug class information and interactions help ensure safer prescribing habits. Additionally, since all providers will be referencing the same list, prescribing errors secondary to ignorance of other medications the patient is on should be decreased.

**Barriers**

No comprehensive solution for solving the problem of the multiple inaccurate outpatient medication lists exits at this time. It is an important problem to tackle, and clearly a difficult one. We acknowledge that the implementation of this framework will require cooperation among several groups, however, we must look towards designing a solution to battle the inefficiency and lack of scalability that is today’s normal. The result of our current system is wasted time reconciling patient lists in addition to the high cost of ADEs. Convincing patients to remain diligent in keeping an up-to-date list is difficult. Physicians are typically slow to adopt new technology and are resistant to tools that ask them to check more boxes. Secure storage and sharing of electronic information is a challenging task. HIEs are local and an organized network to connect them must be considered. Finally, one must demonstrate that cost savings is worth the expenditure of implementation. Each problem deserves attention and must also be considered together when designing a comprehensive solution.

**Medication Harmony**

The triple aim of medication reconciliation is: reasonable time to reconcile, high level of accuracy, and improved patient activation. Instead of each patient and provider struggling to maintain an individual copy of a medication list, the list will be shared among providers and audited by providers and the patient himself. Increased coordination, safety, and education afforded by the centralized list will clearly have the greatest benefit for the patient who verifies and modifies the medication list. However, within the proposed solution, it is important to recognize that active patient involvement is not necessary for maintaining a medication list. Even without active patient involvement, centralization among physicians, hospitals, and pharmacies should be sufficient for obtaining a reasonable proxy to the medication list. In addition, providers can be assured that changes made during a provider-patient interaction will permeate to all other providers. Therefore, instead of each provider having a partial medication list that is often inaccurate, each provider will be able to view the same federated list and can change and prescribe medications more safely. The expectation is that if the task is not overwhelming to the provider, he will be more willing to ensure the accuracy of the list. Finally, a centralized system aids the institutions and individuals attempting to reconcile medications at every transition of care. The process of reconciliation requires significant time and resources. It is a process that is repeated often and at every outpatient visit or transition of care. Tremendous efficiency gains will be realized by implementing the above framework.

**Conclusions**

“You’ve got to be very careful if you don’t know where you are going, because you might not get there.”
- Yogi Berra

Research identifying medication management problems and solutions supporting such a framework has thus far been considered individually. Patients are willing and able to interact with electronic medication tools to manage their health. Medication lists are currently being stored electronically and can be accessed by trusted sources via an HIE. Considerable time and effort is made reconciling the medication list at each transition of care. Inaccuracies in the medication list result in ADEs causing poor patient outcomes and increased cost. We believe that the momentum of the health care system flows towards interoperability, the sharing of data, and the promise of decreased cost. Our lab has started development of this idea into an existing project involving elder care management.

A comprehensive solution involves:
- Providing a pre-populated list with easily accessed and actionable information for patients
- Patients and providers working off the same list
Simplifying and saving time during the process of medication reconciliation  
Encouraging legislation allowing for easier transmission of medical information over trusted networks  
Continued maturity of the FHIR standard and the opening of APIs by EHR vendors  
Financial incentives that continue to emphasize decreasing hospital readmission rates and preventable events

We must utilize the trusted network provided by a HIE and automate the federation of information from multiple sources into a central database. Most importantly, by allowing the patient access to the same list used by providers, patient input into the list will more accurately reflect the daily medications actually taken and allow safer prescribing practices. Imagine a future where patients, doctors, hospitals and pharmacies have easy access to a complete and accurate medication list. We have the potential to drastically reduce adverse drug events both inside and outside the walls of a healthcare setting and should strive to accomplish such goals.

Funding

This work was supported by the Agency for Healthcare Research and Quality grant number R01HS021495 in addition to training grant T15LM007092-23 from the National Library of Medicine.

References

Use of a Patient Portal During Hospital Admissions to Surgical Services

Jamie R. Robinson, M.D., Sharon E. Davis, M.Stat., Robert M. Cronin M.D., M.S., and Gretchen P. Jackson, M.D., Ph.D.
Vanderbilt University Medical Center, Nashville, TN

Abstract

Patient portal research has focused on medical outpatient settings, with little known about portal use during hospitalizations or by surgical patients. We measured portal adoption among patients admitted to surgical services over two years. Surgical services managed 37,025 admissions of 31,310 unique patients. One-fourth of admissions (9,362, 25.3%) involved patients registered for the portal. Registration rates were highest for admissions to laparoscopic/gastrointestinal (55%) and oncology/endocrine (50%) services. Portal use occurred during 1,486 surgical admissions, 4% of all and 16% of those registered at admission. Inpatient portal use was associated with patients who were white, male, and had longer lengths of stay ($p < 0.01$). Viewing health record data and secure messaging were the most commonly used functions, accessed in 4,836 (72.9%) and 1,626 (24.5%) user sessions. Without specific encouragement, hospitalized surgical patients are using our patient portal. The surgical inpatient setting may provide opportunities for patient engagement using patient portals.

Introduction

Patient portals are web-based applications that enable patients to view portions of their electronic health record (EHR) and interact with their healthcare providers. The United States government defines a patient portal as “a secure online website that gives patients convenient 24-hour access to personal health information from anywhere with an Internet connection.” The data within a portal is typically managed by a healthcare institution and allows patients to have access to personal health information, including recent doctor visits, discharge summaries, medications, immunizations, allergies, and laboratory results. More advanced portals enable patients to schedule appointments, message their providers, and sometimes maintain personal health records. Increasingly, health care systems offer portals to their patients, and consumers adopt them quickly. Hospitals are motivated to provide patient portals by financial incentives created by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 and Meaningful Use criteria. Consumers are also demanding such technology and transparency from their health care providers.

The majority of research about patient portals has been performed in the primary care or medical specialty settings with a paucity of research focusing on acute care specialties, such as surgery, or acute care settings, such as the hospital. Two recent systematic reviews of over 100 studies on the effectiveness of patient portals revealed only three studies exploring portal use outside of primary care or medical specialties. Our prior research demonstrated that after broad deployment of a patient portal across clinical specialties, surgeons were the second most frequent specialty to use patient-provider messaging. Further, messaging adoption by surgical patients and providers grew rapidly across surgical subspecialties. As healthcare organizations increasingly deploy patient portals across clinical specialties to meet Meaningful Use objectives, we anticipate the use of patient portals by the understudied acute care population to continue to grow.

Many trials investigating patient portals involve outpatient management of chronic diseases, but very little is known regarding patient use of patient portals while in the hospital for an acute illness or after surgery. Masterson et al have described their development of a personalized inpatient portal to improve patient engagement while in the hospital, but results are not yet reported. Very small studies have assessed the efficacy of providing patients with tablet computers for specific encouragement of portal use during inpatient stay, showing patients utilized and appreciated the ability to view their health information. At Brigham and Women’s Hospital, a web-based patient-centered tool kit offering many common patient portal functions was implemented and evaluated in the medical intensive care unit and oncology unit settings. Studies of this system have shown encouraging trends for the adoption and sustained usage of such technologies in the acute care setting. Although the functions commonly offered by patient portals meet important needs of hospitalized patients, there has been a reluctance to encourage portal usage by hospitalized patients. Many patient portals have policies that would discourage inpatient usage, such
as specific delays for availability of test results or several business day expectations for answering of secure messages. Nonetheless, healthcare consumers often find innovative ways to use health information technologies. To address existing gaps in the literature about the use of patient portals by surgical patients in acute care settings, we sought to characterize the adoption and use of a patient portal by patients admitted to surgical services at a large academic medical center.

Methods

Study Setting
This study was conducted at Vanderbilt University Medical Center (VUMC), a private, non-profit, academic institution in Nashville, Tennessee, which provides primary and regional referral care to adults and children. VUMC encompasses Vanderbilt University Hospital (VUH) and Vanderbilt Children’s Hospital (VCH), with over 900 inpatient beds and 50,000 inpatient admissions per year.

Patient Portal
VUMC launched the My Health At Vanderbilt (MHAV) patient portal in 2005, with implementation across the clinical enterprise completed in 2007. After initial implementation, a physician champion introduced MHAV to providers, and technical support staff was available to patients, physicians, and staff as the portal was introduced in individual clinical units. MHAV was promoted to patients through flyers posted in outpatient clinics. This process was repeated, beginning in adult primary care, and then extending to adult and pediatric specialties. Programs did not promote MHAV use in the inpatient setting until 2014, when a link to the MHAV was made available through interactive television in the hospital rooms of VUH.

All patients who receive medical care at VUMC may register for access to MHAV. MHAV users may authorize another individual, termed a delegate, to access their MHAV account on their behalf. Our policies for MHAV accounts for pediatric patients are similar to those developed for other major children’s hospitals. For patients under 13 years of age, parents or guardians (called surrogates) may access MHAV account on behalf of their child. Adolescents 13 years of age and older may have their own, parent-controlled MHAV accounts. MHAV provides access to selected portions of the medical record, appointment scheduling, account and bill management, targeted health education materials, and secure messaging with healthcare providers. MHAV is now a well-established patient portal, with over 327,000 registered users and over 300,000 logins per month by 50,000 unique users. MHAV is directly linked to the VUMC EHR, StarPanel, and thus, content is continually updated. MHAV allows access to selected health information from the EHR, including clinical visit summaries, laboratory results, and medication lists. Some information is immediately available and other sensitive content is only viewable after short delays to allow for physician review and management. MHAV messages are managed by clinical groups based on provider preferences. Some providers directly answer all patient-initiated messages, and others have messages triaged by administrative and clinical staff members, any of whom may respond. Tailored educational materials are available within MHAV based on patient problems and medication lists. Specific policies and procedures developed to enhance patient and provider adoption are published elsewhere.

Study Population
We examined all admissions to surgical services at VUMC and all use of the MHAV patient portal by patients admitted by a surgical service or their delegates or surrogates between January 1, 2012 and December 31, 2013. This time period was chosen after a rise in anecdotal reporting of inpatient MHAV usage and prior to the promotion of portal registration and usage through interactive television in the hospital. We sought to examine inpatient portal access in the absence of specific programs to promote such usage.

Measures
For each admission during the study period, we recorded patient age, sex, and race, as well as admitting service, International Classification of Disease (ICD-9) admission diagnosis code, and length of stay. We categorized admitting services using 21 surgical specialties reflecting the departmental organization at VUMC: cardiac, thoracic, otolaryngology, emergency general, general, gastrointestinal/laparoscopic, liver transplant, neurosurgery, oncology/endocrinology, oral/maxillofacial, orthopedic, pediatric, pediatric trauma, pediatric urology, plastic, renal transplant, spinal, trauma, burn, urology, and vascular. MHAV users include VUMC patients who have registered for MHAV, delegates, and surrogates. For each admission, we considered the patient registered with MHAV if they had a portal account by the time of discharge. MHAV use during a hospital admission was defined as any MHAV
activity through the admitted patient’s or any affiliated delegate/surrogate accounts occurring between the date/time of admission and the date/time of discharge. We determined the total number of inpatient MHAV user sessions across user types. For each session, we classified the type of portal function utilized as account management, appointments, education materials, laboratory test results, messaging, or other.

Analysis
We calculated the total number of inpatient admissions to each surgical service, as well as the number of these surgical patients who were registered for MHAV either before or during their admission. We also calculated the number of these patients who specifically registered for MHAV during inpatient stay. We constructed descriptive distributions and summary statistics of MHAV registration and use status across patient demographics and admission characteristics. Continuous variables were summarized with medians and inter-quartile ranges. Categorical variables were summarized as counts and frequencies. We modeled inpatient use among admissions of registered patients using a logistic model controlling for month of admission, race, sex, age at admission, length of stay, and admitting service. Standard errors were adjusted to account for correlation among multiple admissions for the same patient. All analyses were conducted in R version 3.0.1.

Results
During the study period, VUMC surgical services managed 37,025 admissions of 31,310 unique patients. Demographics of the unique patients admitted to a surgical service listed in Table 1.

Table 1. Demographic and clinical characteristics of unique patients admitted to a surgical service at VUMC 2012-2013. Counts and percentages or median and IQR.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All</th>
<th>%</th>
<th>VCH %</th>
<th>VUH %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>31,310</td>
<td>5,002</td>
<td>26,308</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>26,380</td>
<td>84.3</td>
<td>3,845</td>
<td>76.9</td>
</tr>
<tr>
<td>Black</td>
<td>3,497</td>
<td>11.2</td>
<td>715</td>
<td>14.3</td>
</tr>
<tr>
<td>Unknown</td>
<td>975</td>
<td>3.1</td>
<td>298</td>
<td>6</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>355</td>
<td>1.1</td>
<td>123</td>
<td>2.5</td>
</tr>
<tr>
<td>Native American/Alaskan</td>
<td>103</td>
<td>0.3</td>
<td>21</td>
<td>0.4</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>17,939</td>
<td>57.3</td>
<td>2,875</td>
<td>57.5</td>
</tr>
<tr>
<td>Female</td>
<td>13,371</td>
<td>42.7</td>
<td>2,127</td>
<td>42.5</td>
</tr>
<tr>
<td>Age at first admission (years)</td>
<td>49 (IQR 26-63)</td>
<td>6 (IQR 2-13)</td>
<td>54 (IQR 39-65)</td>
<td></td>
</tr>
<tr>
<td>Age categories (decades in years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td>3,175</td>
<td>10.1</td>
<td>3,152</td>
<td>63</td>
</tr>
<tr>
<td>10-19</td>
<td>2,694</td>
<td>8.6</td>
<td>1,792</td>
<td>35.8</td>
</tr>
<tr>
<td>20-29</td>
<td>2,863</td>
<td>9.1</td>
<td>39</td>
<td>0.8</td>
</tr>
<tr>
<td>30-39</td>
<td>3,099</td>
<td>9.9</td>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>40-49</td>
<td>4,040</td>
<td>12.9</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>50-59</td>
<td>5,457</td>
<td>17.4</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>60-69</td>
<td>5,609</td>
<td>17.9</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>70-79</td>
<td>3,218</td>
<td>10.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>80-89</td>
<td>1,035</td>
<td>3.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90 or older</td>
<td>120</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age categories (pediatric)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1yr</td>
<td>1156</td>
<td>3.7</td>
<td>1145</td>
<td>22.9</td>
</tr>
<tr>
<td>2-5yrs</td>
<td>1,142</td>
<td>3.6</td>
<td>1,132</td>
<td>22.6</td>
</tr>
<tr>
<td>6-10yrs</td>
<td>1,072</td>
<td>3.4</td>
<td>1,070</td>
<td>21.4</td>
</tr>
<tr>
<td>11-15yrs</td>
<td>1,231</td>
<td>3.9</td>
<td>1,208</td>
<td>24.2</td>
</tr>
<tr>
<td>16-18yrs</td>
<td>928</td>
<td>3</td>
<td>364</td>
<td>7.3</td>
</tr>
<tr>
<td>over 18yrs</td>
<td>25,781</td>
<td>82.3</td>
<td>83</td>
<td>1.7</td>
</tr>
</tbody>
</table>
**MHAV Registration Status**

Of the 37,025 admissions during the study period, 9,362 (25.3%) involved patients registered for MHAV and 7,549 (24.1%) unique patients were registered for MHAV during at least one admission in the study period. In 194 admissions, the patient registered for MHAV during an inpatient stay rather than enrolling in MHAV prior to the admission. The MHAV registration rate was higher at VUH than VCH, with 27.0% of unique patients admitted to VUH having a portal account compared to 8.8% among unique patients admitted to VCH.

Table 2 presents the demographics for all patients admitted to a surgical service compared to those registered for MHAV. Patients registered for MHAV differed from the entire patient cohort on each demographic characteristic, both overall and within each hospital ($p < 0.01$) with the exception of sex among patients admitted to VCH ($p = 0.29$). White and Asian/Pacific Islander patients were more likely to have a MHAV account than were Black, Native American/Alaskan, and other/unreported race patients. Overall, patients in their 50s and 60s were most likely to be registered for MHAV. Among patients admitted to VUH, female patients were more likely to be registered for MHAV compared to male patients (33.3% vs. 22.3%, respectively).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All (n)</th>
<th>Registered (n)</th>
<th>Registered (% of total)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>26,380</td>
<td>6,812</td>
<td>25.8</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>355</td>
<td>88</td>
<td>24.8</td>
</tr>
<tr>
<td>Native American/Alaskan</td>
<td>103</td>
<td>18</td>
<td>17.5</td>
</tr>
<tr>
<td>Black</td>
<td>3,497</td>
<td>545</td>
<td>15.6</td>
</tr>
<tr>
<td>Unknown</td>
<td>975</td>
<td>86</td>
<td>8.8</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>13,371</td>
<td>3,941</td>
<td>29.5</td>
</tr>
<tr>
<td>Male</td>
<td>17,939</td>
<td>3,608</td>
<td>20.1</td>
</tr>
<tr>
<td><strong>Age at first admission (years)</strong></td>
<td>49 (IQR 26-73)</td>
<td>54 (IQR 41-64)</td>
<td></td>
</tr>
<tr>
<td><strong>Age categories (decades)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 10</td>
<td>3,175</td>
<td>277</td>
<td>8.7</td>
</tr>
<tr>
<td>10-19</td>
<td>2,694</td>
<td>202</td>
<td>7.5</td>
</tr>
<tr>
<td>20-29</td>
<td>2,863</td>
<td>452</td>
<td>15.8</td>
</tr>
<tr>
<td>30-39</td>
<td>3,099</td>
<td>834</td>
<td>26.9</td>
</tr>
<tr>
<td>40-49</td>
<td>4,040</td>
<td>1,193</td>
<td>29.5</td>
</tr>
<tr>
<td>50-59</td>
<td>5,457</td>
<td>1,768</td>
<td>32.4</td>
</tr>
<tr>
<td>60-69</td>
<td>5,609</td>
<td>1,821</td>
<td>32.5</td>
</tr>
<tr>
<td>70-79</td>
<td>3,218</td>
<td>801</td>
<td>24.9</td>
</tr>
<tr>
<td>80-89</td>
<td>1,035</td>
<td>186</td>
<td>18</td>
</tr>
<tr>
<td>90 or older</td>
<td>120</td>
<td>15</td>
<td>12.5</td>
</tr>
<tr>
<td><strong>Age categories (pediatric)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 6 months</td>
<td>542</td>
<td>34</td>
<td>6.3</td>
</tr>
<tr>
<td>6-12 months</td>
<td>248</td>
<td>33</td>
<td>13.3</td>
</tr>
<tr>
<td>12-24 months</td>
<td>366</td>
<td>50</td>
<td>13.7</td>
</tr>
<tr>
<td>2-5yrs</td>
<td>1,142</td>
<td>96</td>
<td>8.4</td>
</tr>
<tr>
<td>6-10yrs</td>
<td>1,072</td>
<td>75</td>
<td>7.0</td>
</tr>
<tr>
<td>11-15yrs</td>
<td>1,231</td>
<td>91</td>
<td>7.4</td>
</tr>
<tr>
<td>16-18yrs</td>
<td>928</td>
<td>69</td>
<td>7.4</td>
</tr>
<tr>
<td>over 18yrs</td>
<td>25,781</td>
<td>7,101</td>
<td>27.5</td>
</tr>
</tbody>
</table>

The number and proportion of patients registered for MHAV by surgical admitting service are presented in Table 3. At the adult hospital, VUH, 8,851 of 31,448 (28.1%) admissions to surgical services involved patients registered for MHAV compared to 511 of 5,577 surgical admissions (9.2%) at VCH. The surgical services with the highest rate of MHAV registration were adult gastrointestinal/laparoscopic (54.5%) and adult oncology/endocrinology (49.6%). The surgical
services with the lowest rates of MHAV registration were pediatric trauma (1.2%) and burn (1.8%).

Table 3. Surgical admitting service and MHAV registration among 2012-2013 VUMC admissions, categorized into Vanderbilt Children’s Hospital (VCH) and Vanderbilt University Hospital (VUH). (Reg = Registered)

<table>
<thead>
<tr>
<th>Surgical Admitting Service</th>
<th>All (n)</th>
<th>Reg (n)</th>
<th>Reg (%)</th>
<th>All (VCH)</th>
<th>Reg(n) (VCH)</th>
<th>Reg(%) (VCH)</th>
<th>All (VUH)</th>
<th>Reg(n) (VUH)</th>
<th>Reg(%) (VUH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>37,025</td>
<td>9,362</td>
<td>25.3</td>
<td>5,577</td>
<td>511</td>
<td>9.2</td>
<td>31,448</td>
<td>8,851</td>
<td>28.1</td>
</tr>
<tr>
<td>Year of admission</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>18,270</td>
<td>4,471</td>
<td>24.5</td>
<td>2,655</td>
<td>215</td>
<td>8.1</td>
<td>15,615</td>
<td>4,256</td>
<td>27.3</td>
</tr>
<tr>
<td>2013</td>
<td>18,755</td>
<td>4,891</td>
<td>26.1</td>
<td>2,922</td>
<td>296</td>
<td>10.1</td>
<td>15,833</td>
<td>4,595</td>
<td>29</td>
</tr>
<tr>
<td>GI/Laparoscopic</td>
<td>1,574</td>
<td>858</td>
<td>54.5</td>
<td>1</td>
<td>100</td>
<td>0.0</td>
<td>1,573</td>
<td>857</td>
<td>54.5</td>
</tr>
<tr>
<td>Oncology/Endocrine</td>
<td>1,206</td>
<td>598</td>
<td>49.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1,206</td>
<td>598</td>
<td>49.6</td>
</tr>
<tr>
<td>Spinal</td>
<td>27</td>
<td>13</td>
<td>48.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>13</td>
<td>48.1</td>
</tr>
<tr>
<td>Thoracic</td>
<td>1,068</td>
<td>481</td>
<td>45.0</td>
<td>1</td>
<td>100</td>
<td>0.0</td>
<td>1,067</td>
<td>480</td>
<td>45</td>
</tr>
<tr>
<td>General</td>
<td>2,006</td>
<td>855</td>
<td>42.6</td>
<td>23</td>
<td>1</td>
<td>4.3</td>
<td>1,983</td>
<td>854</td>
<td>43.1</td>
</tr>
<tr>
<td>Renal Transplant</td>
<td>504</td>
<td>205</td>
<td>40.7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>500</td>
<td>205</td>
<td>41</td>
</tr>
<tr>
<td>Liver Transplant</td>
<td>534</td>
<td>207</td>
<td>38.8</td>
<td>3</td>
<td>33.3</td>
<td>0.0</td>
<td>531</td>
<td>206</td>
<td>38.8</td>
</tr>
<tr>
<td>Neurological</td>
<td>4,239</td>
<td>1,461</td>
<td>34.5</td>
<td>391</td>
<td>84</td>
<td>21.5</td>
<td>3,848</td>
<td>1,377</td>
<td>35.8</td>
</tr>
<tr>
<td>Urology</td>
<td>2,658</td>
<td>805</td>
<td>30.3</td>
<td>50</td>
<td>4</td>
<td>8</td>
<td>2,608</td>
<td>801</td>
<td>30.7</td>
</tr>
<tr>
<td>Cardiac</td>
<td>957</td>
<td>279</td>
<td>29.2</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>948</td>
<td>279</td>
<td>29.4</td>
</tr>
<tr>
<td>Emergency General</td>
<td>1,444</td>
<td>385</td>
<td>26.7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1,443</td>
<td>385</td>
<td>26.7</td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>2,030</td>
<td>521</td>
<td>25.7</td>
<td>574</td>
<td>75</td>
<td>13.1</td>
<td>1,456</td>
<td>446</td>
<td>30.6</td>
</tr>
<tr>
<td>Orthopedic/Rehab</td>
<td>6,602</td>
<td>1,672</td>
<td>25.3</td>
<td>1,011</td>
<td>84</td>
<td>8.3</td>
<td>5,591</td>
<td>1,588</td>
<td>28.4</td>
</tr>
<tr>
<td>Vascular</td>
<td>525</td>
<td>121</td>
<td>23.1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>524</td>
<td>121</td>
<td>23.1</td>
</tr>
<tr>
<td>Plastic</td>
<td>1,435</td>
<td>283</td>
<td>19.7</td>
<td>451</td>
<td>53</td>
<td>11.8</td>
<td>984</td>
<td>230</td>
<td>23.4</td>
</tr>
<tr>
<td>Oral/Maxillofacial</td>
<td>286</td>
<td>44</td>
<td>15.4</td>
<td>49</td>
<td>2</td>
<td>4.1</td>
<td>237</td>
<td>42</td>
<td>17.7</td>
</tr>
<tr>
<td>Pediatric Urology</td>
<td>269</td>
<td>25</td>
<td>9.3</td>
<td>269</td>
<td>25</td>
<td>9.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pediatric</td>
<td>2,292</td>
<td>175</td>
<td>7.6</td>
<td>2,287</td>
<td>175</td>
<td>7.7</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Trauma</td>
<td>6,032</td>
<td>352</td>
<td>5.8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>6,022</td>
<td>352</td>
<td>5.8</td>
</tr>
<tr>
<td>Burn</td>
<td>1,083</td>
<td>19</td>
<td>1.8</td>
<td>191</td>
<td>2</td>
<td>1</td>
<td>892</td>
<td>17</td>
<td>1.9</td>
</tr>
<tr>
<td>Pediatric Trauma</td>
<td>254</td>
<td>3</td>
<td>1.2</td>
<td>251</td>
<td>3</td>
<td>1.2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Inpatient Use of MHAV

Portal usage occurred during 1,486 surgical admissions (4% of all admissions and 16% of registered user admissions) involving 1,270 unique patients. 6,634 portal user sessions occurred during surgical inpatient admissions. For admissions during which MHAV was accessed, the median number of MHAV sessions was 2 (IQR 1-4); however, during some admissions, patients accessed MHAV more than 20 times, with a few users accessing MHAV over 80 times during admission. Normalizing by length of stay, the median number of MHAV sessions per inpatient day was 2.0 (IQR 1.0-3.1) among admissions with MHAV use.

In unadjusted tests among admissions involving MHAV registered patients, admissions with inpatient portal use differed from those without portal use in terms of length of stay, race, sex, and admitting service (p < 0.01), but did not differ on patient age. These findings were observed overall and within VUH admissions only. Among admissions to VCH, admissions with portal use were longer than admissions without use (median LOS 5 vs. 3 days; p < 0.01).

Adjusting for patient demographics and admission characteristics, white race, male sex, increased length of hospital stay, and admitting service were associated with inpatient portal use (p < 0.01). Figure 1 demonstrates the odds ratios (OR) for inpatient portal use based on demographics and admission service. Black patients were significantly less likely than white patients to use the portal or have MHAV accessed on their behalf during hospitalization (OR 0.53, 95% CI 0.39-0.71). Male patients were more likely to use the portal or have MHAV accessed on their behalf during hospitalization than female patients (OR 1.33, 95% CI 1.17-1.52). Compared to general surgery admissions, admissions to the liver transplant service were at 76% higher odds of portal use (OR 1.76, 95% CI 1.19-2.62). Admissions to the liver transplant service were also more likely to use the portal than those to neurological, plastic,
gastrointestinal/laparoscopic, otolaryngology, and orthopedic surgery services. Otolaryngology (OR 0.69, 95% CI 0.48-0.98), gastrointestinal/laparoscopic (OR 0.57, 95% CI 0.41-0.80), and orthopedic surgery (OR 0.73, 95% CI 0.56-0.94) admissions showed a decreased likelihood of portal use compared to general surgery admissions.

Among admissions with inpatient use of MHAV, the portal was accessed through the patient’s account in 92.7% of admissions, through a delegate account in 2.6% of admissions, and through a surrogate account in 5.5% of admissions (see Table 4). Although patients utilized a variety of portal functions, viewing health record data (i.e. laboratory results, medication lists, or clinical documents) and secure patient-provider messaging were the most common, accessed in 4,836 (72.9%) and 1,626 (24.5%) of total inpatient user sessions, respectively (see Table 4).

Table 4: Number of inpatient user sessions accessing each MHAV function, overall and by user role.

<table>
<thead>
<tr>
<th>Function</th>
<th>Any user (n)</th>
<th>Patient (n)</th>
<th>Delegate (n)</th>
<th>Surrogate (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sessions</td>
<td>6,634</td>
<td>6,243</td>
<td>127</td>
<td>264</td>
</tr>
<tr>
<td>Viewing health record</td>
<td>4,836</td>
<td>4,563</td>
<td>77</td>
<td>196</td>
</tr>
<tr>
<td>Messaging</td>
<td>1,626</td>
<td>1,489</td>
<td>54</td>
<td>83</td>
</tr>
<tr>
<td>Educational materials</td>
<td>521</td>
<td>521</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Appointments</td>
<td>495</td>
<td>462</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>Account management</td>
<td>72</td>
<td>67</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>120</td>
<td>112</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

Tables 5 and 6 describe the most prevalent ICD-9 diagnosis codes for all admissions and admissions with inpatient use of the patient portal to VUH (adults) and VCH (pediatrics), respectively. The three most frequent diagnoses among adult patients who utilized the portal while hospitalized were postoperative infection, morbid obesity, and intestinal obstruction. In contrast, the 3 most frequent diagnoses among pediatric patients who used the portal while inpatient were scoliosis and kyphoscoliosis, esophageal reflux, and hypertrophy of tonsils with adenoids.
was liver transplantation, potentially due to frequent laboratory monitoring and clinical complexity of patients.

Table 5: Top 10 most prevalent ICD9 diagnosis codes of patients at VUH (adults; n = # of admissions)

<table>
<thead>
<tr>
<th>ICD-9</th>
<th>n</th>
<th>Admissions to VUH with inpatient MHAV use</th>
</tr>
</thead>
<tbody>
<tr>
<td>185 - Mal neoplasm prostate</td>
<td>771</td>
<td>998.59 - Other postop infection</td>
</tr>
<tr>
<td>278.01 - Morbid obesity</td>
<td>759</td>
<td>278.01 - Morbid obesity</td>
</tr>
<tr>
<td>998.59 - Other postop infection</td>
<td>610</td>
<td>560.9 - Intestinal obstruction</td>
</tr>
<tr>
<td>189.0 - Mal neoplasm kidney</td>
<td>464</td>
<td>189.0 - Mal neoplasm kidney</td>
</tr>
<tr>
<td>715.36 - Osteoarthrosis lower leg</td>
<td>433</td>
<td>403.9 - Hypertensive chronic kidney disease</td>
</tr>
<tr>
<td>715.35 - Osteoarthrosis pelvis</td>
<td>254</td>
<td>198.3 - Secondary mal neoplasm brain/spinal</td>
</tr>
<tr>
<td>560.9 - Intestinal obstruction</td>
<td>242</td>
<td>V55.2 - Attention to ileostomy</td>
</tr>
<tr>
<td>733.82 - Nonunion of fracture</td>
<td>223</td>
<td>715.36 - Osteoarthrosis lower leg</td>
</tr>
<tr>
<td>403.91 - Hypertensive chronic kidney disease</td>
<td>221</td>
<td>562.11 - Diverticulitis of colon</td>
</tr>
<tr>
<td>414.01 - Coronary Atherosclerosis</td>
<td>194</td>
<td>997.49 - Other digestive sys complications</td>
</tr>
</tbody>
</table>

Table 6: Top 10 most prevalent ICD9 diagnosis codes of patients at VCH (children; n = # of admissions)

<table>
<thead>
<tr>
<th>ICD-9</th>
<th>n</th>
<th>Admissions to VUH with inpatient MHAV use</th>
</tr>
</thead>
<tbody>
<tr>
<td>750.5 - Hypertrophic pyloric stenosis</td>
<td>233</td>
<td>737.30 - Scoliosis and kyphoscoliosis</td>
</tr>
<tr>
<td>540.9 - Acute appendicitis w/o peritonitis</td>
<td>224</td>
<td>530.81 - Esophageal reflux</td>
</tr>
<tr>
<td>474.10 - Hypertrophy of tonsil w/ adenoids</td>
<td>145</td>
<td>474.10 - Hypertrophy of tonsil w/ adenoids</td>
</tr>
<tr>
<td>737.30 - Scoliosis and kyphoscoliosis</td>
<td>134</td>
<td>276.51 - Dehydration</td>
</tr>
<tr>
<td>540.0 - Acute appendicitis w/ peritonitis</td>
<td>126</td>
<td>996.63 - Complication nervous sys implant</td>
</tr>
<tr>
<td>812.41 - Supracondylar fracture humerus</td>
<td>109</td>
<td>787.22 - Dysphagia oropharyngeal phase</td>
</tr>
<tr>
<td>998.11 - Hemorrhage complicating procedure</td>
<td>89</td>
<td>560.81 - Peritoneal adhesions w/ obstruction</td>
</tr>
<tr>
<td>756.0 - Congenital anomalies of skull/face</td>
<td>86</td>
<td>556.9 - Ulcerative colitis unspecified</td>
</tr>
<tr>
<td>996.2 - Complication nervous sys implant</td>
<td>85</td>
<td>756.19 - Other congenital anomalies spine</td>
</tr>
<tr>
<td>540.1 - Acute appendicitis w/ abscess</td>
<td>84</td>
<td>742.59 - Other cong anomalies spinal cord</td>
</tr>
</tbody>
</table>

Discussion

This study documents modest and somewhat unexpected usage of a patient portal by hospitalized surgical patients; it is one of the first studies to report inpatient portal adoption outside of a specific program or technology designed for the hospital setting. Without promotion for use in the inpatient setting, 4% of all admitted surgical patients and 16% of patients registered for the portal utilized the portal while in the hospital. With a known lack of research about technologies to engage patients in the inpatient setting, this study suggests that existing technologies such as patient portals may have a role in meeting the needs of hospitalized patients and their families.

Our study showed that patient portals were more likely to be used during hospitalization for patients who were white, male, and had extended lengths of stay. Outpatient studies of patient portals have shown similar disparities with decreased use by minorities, especially African Americans. In contrast to our findings, prior studies suggest that portal use is fairly similar between women and men, with most studies demonstrating slightly higher registration rates and usage by women. Of note, we cannot determine from usage logs whether the portal was actually used personally by the patient, or rather another individual using the patient’s login information.

In our study, the services with the most registered portal users included those with significant pre-operative relationships, including gastrointestinal and laparoscopic (including a large majority of bariatric surgery patients) and oncology and endocrine surgery. ICD-9 diagnosis codes for adults registered for the portal aligned with the service designations, with the most common diagnoses being prostate cancer and morbid obesity. Interestingly, the most frequent ICD-9 code of adult patients using the portal while inpatient was post-operative infection, suggesting patients with a complication may be more likely to utilize the portal to view personal health information and contact providers. The only service that showed portal usage increased over that of general surgery in hospitalized patients was liver transplantation, potentially due to frequent laboratory monitoring and clinical complexity of patients.
Prior research has shown encouraging adoption of similar technologies during hospitalization, but usually in the context of a specific research program in which registration was encouraged and usage was supported by training. Wilcox and colleagues piloted a customized inpatient personal health record in cardiothoracic surgery patients and found medication tracking tools to be an effective means to increase inpatient engagement. Burke reported enthusiastic adoption of a web-based multimedia EHR for patients with congenital cardiac disease and their parents with a 93% adoption rate and 67% of use occurring during hospitalization. Notably, this study was conducted in families with children undergoing surgical repair of congenital cardiac abnormalities, who likely have long-standing relationships with their surgeons. O’Leary and colleagues showed that patient use of a portal designed specifically with inpatient information including team members, medication lists, and daily agendas on tablet computers within a general medical service unit could improve the ability of patients to identify physicians and roles by over 25%. In contrast to prior work focused on technologies developed for inpatient setting, our study demonstrated substantial use of a patient portal designed for the outpatient setting, by patients who were hospitalized and their caregivers, without specific encouragement or training, and in the presence of policies that might discourage inpatient use.

There are many potential benefits to using a patient portal during inpatient admissions. First and foremost, even minor surgeries are considered major life events for most patients and families, and they offer “teachable moments” when otherwise unengaged individuals might consider making important healthcare changes. Introducing patient portals during hospitalizations may provide tools for patients and families to learn about health problems and engage in their care. Furthermore, hospital team members, including physicians, residents, and nursing staff are highly dynamic, and many hospitalized patients are unable to identify their physicians. Hospitalized patients frequently have multiple active conditions, tests, and procedures, with acute illness and its associated stress making it difficult for patients and families to retain information provided on daily rounds or at discharge. Others have shown that patients and caregivers desire access to the daily plan of care and team member roles, often not present in patient portals. Patient portals can allow patients to review their health data, schedule and view post-operative appointments, and communicate with providers. In the inpatient setting, hospital staff can provide training and support to assist patients and their families with registration and navigation of portal functions, giving them the knowledge and experience needed to promote ongoing engagement. Use and familiarity with the portal prior to discharge may increase the portal usage on an outpatient basis. For example, patients may feel more comfortable communicating problems or concerns post-operatively through secure messaging after using it as an inpatient.

Such changes could have a significant impact on surgical workflow. Some patients who undergo certain operations may not require a face-to-face follow up, and provider-patient messaging could be utilized to ensure the patient is recovering as expected post-operatively. A pilot study at our institution has shown that over three-fourths of patients undergoing elective general surgery procedures were satisfied with online follow up, and post-operative complications were not missed by online visits. Portal follow-up can potentially prevent patients from travelling long distances or missing work or school for unnecessary face-to-face clinic visits. This approach also benefits providers as follow-up appointments are typically included within the global payment period.

This study has important limitations. The design is retrospective, and the research was done at single large academic medical center with a locally-developed patient portal. The findings may not apply to other clinical settings or portal implementations, and therefore may not be generalizable to all hospitals. However, our portal functions and many of the MHAV procedures and policies are similar to those reported by others. One main difference is that MHAV was broadly deployed across clinical specialties soon after implementation, and our findings represent those of an established portal in use across the clinical enterprise for over 5 years. We have not assessed factors that may contribute to adoption and usage of the portal, such as encouragement by specific providers or teams. Further, we do not know the platform on which the portal was used by patients, the clinical context in which the portal was accessed, or other measures of usability or satisfaction, which would further inform the interpretation of our usage data. These questions are the subject of our ongoing research projects.

**Conclusions**

This study demonstrates modest use of a patient portal by hospitalized surgical patients without specific encouragement. Disparities in portal adoption among minority patients may occur in the inpatient setting. Although designed for the outpatient setting, patient portals may have a role in meeting consumer health information needs and engaging surgical patients both during and after hospitalizations. The perioperative period may offer a uniquely teachable time in which to engage patients and families in their care, and using a portal during hospitalization could...
support online postoperative follow up, which can benefit both patients and providers. Additional research is needed to determine the best ways to leverage patient portals during inpatient admissions to improve care.

Acknowledgements

Jamie Robinson and Sharon Davis were supported by the 5T15LM007450-12 training grant from the National Library of Medicine.

References


1976
Characterization of Temporal Semantic Shifts of Peer-to-Peer Communication in a Health-Related Online Community: Implications for Data-driven Health Promotion

Vishnupriya Sridharan, B.Tech1, Trevor Cohen MBChB, PhD1, Nathan Cobb, MD2, Sahiti Myneni, PhD, MSE1

1The University of Texas School of Biomedical Informatics at Houston, TX, USA
2Georgetown University Medical Center, Washington, DC, United States

Abstract:

With online social platforms gaining popularity as venues of behavior change, it is important to understand the ways in which these platforms facilitate peer interactions. In this paper, we characterize temporal trends in user communication through mapping of theoretically-linked semantic content. We used qualitative coding and automated text analysis to assign theoretical techniques to peer interactions in an online community for smoking cessation, subsequently facilitating temporal visualization of the observed techniques. Results indicate manifestation of several behavior change techniques such as ‘feedback and monitoring’ and ‘rewards’. Automated methods yielded reasonable results (F-measure=0.77). Temporal trends among relapers revealed reduction in communication after a relapse event. This social withdrawal may be attributed to failure guilt after the relapse. Results indicate significant change in thematic categories such as ‘social support’, ‘natural consequences’, and ‘comparison of outcomes’ pre and post relapse. Implications for development of behavioral support technologies that promote long-term abstinence are discussed.

Introduction and Background:

According to CDC statistics of 2015, chronic diseases – such as hypertension, stroke, cancer and diabetes, are responsible for 7 of 10 deaths every year in the United States and treating people with chronic diseases accounts for 86% of all healthcare costs [1]. These chronic conditions are often caused by health-related behaviors such as tobacco smoking, which is “the leading cause of preventable disease and death in the United States, resulting in approximately 480,000 premature deaths and more than $300 billion in direct health care expenditures and productivity losses each year” [2]. Health promotion campaigns emphasize the need to create behavior change avenues through theoretically designed interventions to help people modify this risky health behavior and stay abstinent. Current trends show that online social communities are gaining popularity as behavior modification venues as users of these platforms reach out to their peers and experts for support and guidance irrespective of geographical and demographic boundaries [3]. End-users of these virtual support groups (patients or population at large) have expressed that they trust and rely on the information in social media, and even base their decisions on this information [4]. In addition, using social media peer interactions for behavior analysis has great potential because (a) unlike conventional survey-based or controlled laboratory studies, content on social media is spontaneous and unprompted [3], and (b) electronically captured communication is amenable to large scale text analysis thereby scaling up traditional socio-behavioral analytical methods to social media platforms in the digital era.

Previous research in this field has focused on analysis of the structural characteristics of the online communities [5], development of theory-guided interventions [6], and validation of social support in online communities for behavior change and chronic illness management [7]. Recent studies on social media analysis for smoking cessation have suggested new methodological advances for enhancing user engagement through content-mediated network modeling of peer interactions [8, 9]. Methods of distributional semantics, which learn the relatedness between terms from large electronic text collections, have been used in conjunction with social influence models to characterize content-specific communication patterns underlying behavior change in smoking cessation [10]. Such automated text analysis methods have been used for large scale analysis of content on social networking sites [11-13]. Latent Semantic Analysis (LSA) [14] has been used in conjunction with machine learning techniques to annotate large amount of health information on social media [15-17]. However, computational overhead may pose a problem while using LSA. More scalable distributional semantic methods such as Reflective Random Indexing (RRI) [18] can also learn semantic relatedness from text corpora, including meaningful relations between terms that do not occur together. Semantic space models have been developed to represent text characterizing a specific user’s temporal transitions in online communities [19-21]. Further development of automated methods that enable understanding of individual and network level trends may
provide insight into social influence mechanisms in the context of sustained behavior change. Once analyzed, messages exchanged between users provide a semantic context for their health-related behavior. We posit that the relationship between this semantic context, behavior change events, and user engagement levels can be understood by modeling user communication over time.

Advances in text analytics can enable us to characterize temporal trends in semantic context underlying peer interactions [21-23]. While the overarching goals of our work are to develop informatics-driven methods to develop resource optimized analytics that account for granularity and scalability, and inform the design of consumer-facing digital health platforms for sustained user engagement and long term behavior change, the specific methodological objectives of this paper are three-fold: (a) to employ a behavior change taxonomy for annotation of peer interactions in QuitNet, an online community for smoking cessation, (b) to utilize automated text analysis to scale the results of qualitative analysis to a large dataset, and (c) to visualize peer interactions in QuitNet and characterize user semantics over time, and across smoking and abstinence behaviors.

Materials and Methods:

QuitNet is one of the first online social networks for health behavior change and has been in continuous existence for the past 16 years [24]. Forum interactions were the primary mode of communication and each forum message has a message id, a thread id (the thread in which the message was exchanged), a sender id and recipient id. For the purpose of this study, we considered two subsets of QuitNet data: 16,492 forum messages exchanged between March-April 2007 (we refer to this as “Dataset 1” from now on) and 65,910 forum messages exchanged from January -December 2014 (we refer to this as “Dataset 2” from now on). Overall, there are 82,402 messages and 2,354 unique users who have exchanged messages during these two time periods. The research reported in this manuscript has been reviewed and exempted by the Institutional Review Board at the University of Texas Health Science Center at Houston.

![Figure 1. Methodological outline: Qualitative-Automated-Temporal Analysis](image)

Figure 1. Methodological outline: Qualitative-Automated-Temporal Analysis

Qualitative coding, automated text analysis, and temporal visualization were integrated to conduct the study described in this paper. Figure1 represents the overall methodological details of the study. The first step in this study was the qualitative analysis, guided by an established behavior change taxonomy [25]. 1000 randomly selected messages from the QuitNet Dataset 1 were manually coded into 16 themes to characterize 93 theory-linked behavior change techniques outlined in the taxonomy. In order to validate the generalizability of thematic categorization, scalability of automated methods, and observed semantic shifts before and after relapse in temporal models, we extended our analysis to Dataset 2. The qualitative codes assigned to the subset of 1000 messages in Dataset 1 were then used to annotate the rest of messages in Dataset 1 and entire Dataset 2 using methods from distributional semantics as described in the next section. We then utilized temporal modeling to visualize changes in semantic context over these two time periods (2007 and 2014). On account of its clinical importance, temporal modeling was focused on relapse behavior (self-reported change from ex-smoker to active smoker) during the study period.
**Qualitative methods:**
We selected 1000 messages randomly from the Dataset 1 using a random number generator. Each of these messages was coded into thematic categories by two coders independently to ensure objectivity in the coding process. The themes to which each of the messages was assigned were obtained from the taxonomy of behavior change techniques, which was developed by a large panel of behavior change experts [25]. This taxonomy has 16 thematic categories drawn together from multiple behavior change theories [26] such as Social Change Theory, Social Cognitive Theory, the Health Belief Model, and the Integrative Model of Factors Influencing Smoking. As suggested by Michie at al., the messages were coded to appropriate themes that were explicitly linked to the target behavior (in our case smoking) and target population (QuitNet users). As evident from the sample messages and thematic definitions, a single message may be assigned to multiple taxonomy themes. The definition of each theme and the subcategories of each theme can be found in [25].

**Automated Analysis:**
Methods from distributional semantics in conjunction with a machine learning classifier were used as part of this approach. Incorporation of background semantic information facilitates derivation of meaningful interpretations of QuitNet vector representations, on account of their short and terse textual features [10, 27, 28]. To this end, we used the distributional information from the Touchstone Applied Science Associated (TASA) corpus [29] to provide sufficient semantic context. The TASA is a collection of 44,700 articles that contain 10 million words of unmarked high-school level English text on arts, health, home economics, industrial arts, science, social studies and business. We applied RRI, a variant of Random Indexing which was developed to recognize meaningful relationships between terms without requiring they co-occur directly. This method was applied using Semantic Vectors, an open source package for applying distributional semantics [30]. A stopword list was used in this process [31]. This list contains words that are frequently used in texts but offer little semantic context. We used a dimensionality of 500, with minimum term frequency of 10, maximum term frequency of 15000 and a ‘logentropy’ termweight. We applied RRI to the TASA corpus to obtain *TASA term vectors* - representation of terms in the TASA Corpus. We generated *TASA based QuitNet message vectors* by generating vector representations for each QuitNet message as the sum of the TASA term vectors of the terms it contains, with subsequent normalization. Similarly, we obtained *QuitNet term vectors* by adding the QuitNet message vectors for each term occurring in QuitNet. This step leads to meaningful vector representations for terms that occur in QuitNet, but not in TASA, such as neologisms developed by the community. Finally, we derived a second message vector for each QuitNet message by adding the QuitNet term vectors for the terms it contains, and normalizing the resultant vector to generate *QuitNet message vectors*. This procedure is illustrated in Figure 2.

![Figure 2: Vector generation sequence](image)

From the QuitNet message vectors thus obtained, essential features were selected and extracted for machine learning techniques using the open source Weka package [32]. Each of these vector’s 500 features was considered as individual attributes for the machine learning classification. Each of the themes was considered as a target for classification. Multi label classification was achieved by constructing individual binary classifiers for each theme. The classifier we used for this purpose was the J48 tree which is an improved version of the C4.5 tree [33, 34]. The classifier was trained using ten-fold cross-validation on the QuitNet message vectors representing the training set (manually coded 1000 messages). The trained J48 model was then used to classify QuitNet message vector representations of the entire sets.
of messages Datasets 1 and 2. A random sample of 100 of these messages was used to test for reliability and accuracy of the machine coded messages. These messages were then coded manually by two independent coders and compared with the machine coding results to assess reliability measures. The inter-rater reliability for the machine versus the manual coding was calculated using Cohen’s Kappa measure.

Temporal Modeling:
The communication pattern underlying peer interactions was modeled as a function of time by analyzing the proportion of messages that were categorized as belonging to each thematic category of the BCT taxonomy. The annotated messages from the automated analysis were used for this purpose. A specific category of QuitNet users, ‘relapsers’ (status change from ex-smoker to active smoker as self-reported by the users during study period) was considered for detailed temporal analysis, to characterize the frequency and thematic attributes of communication before and after relapse. The period of time modeled was the period of study (Mar-Apr 2007 & Jan-Dec 2014), i.e, the relapse was used an event marker and the entire set of messages exchanged by a given user before and after their relapse event in the study time period was taken into consideration.

Results and Discussion:
Over 72% users who exchanged messages during the study period were female, and the average age of the users at the time of registration was 45.

Qualitative analysis:
From the manual coding of the 1000 messages, the results indicated that ‘feedback and monitoring’ was the most commonly found theme. The second and third most frequently communicated themes were ‘natural consequences’ and ‘social support’ respectively. The Cohen’s kappa measure between the two coders for the 1000 messages was 0.74. Most differences were related to coding of a single message into multiple themes and were all resolved by discussions. Messages which had users taking “pledges” to not smoke on that day were classified into ‘goals and planning’. Messages where users mention how much time and money they have saved by quitting were classified into the categories ‘feedback and monitoring’ and ‘comparison of outcomes’. QuitNet users also have traditions like “bonfire” where each user virtually throws unused cigarettes into the fire. They spell out the number of cigarettes they are throwing into the fire, thereby, monitoring their own progress. Table 1 shows sample messages for each thematic category. Since communication on social media is unprompted and uninhibited, QuitNet users discuss several extraneous aspects that are not specifically related to smoking. These messages were classified as ‘miscellaneous’ and not considered for further analysis since the taxonomy specifically focuses on target behavior, which in our case is smoking. However, it is important to note that these miscellaneous communications can aid in formation of social bonds, trust, and peer respect which are vital to long-term sustenance of user engagement in these online platforms. In addition, we observed that certain themes did not have ample representation. For instance, the theme ‘associations’, ‘regulation’ and ‘antecedents’ had only four samples out of the 1000 messages. In summary, the mapping of messages to the taxonomy of theoretically-linked behavior change techniques helped us understand if and how such techniques manifest in online platforms that promote health behavior changes. Describing the current landscape of behavioral techniques that manifest spontaneously in peer interactions could help us design better technology platforms that facilitate user interactions resembling other behavior change techniques, to enable users stay quit. For example, content recommendations to QuitNet users providing advice to peers about novel stress management strategies, would integrate ‘shaping knowledge’ techniques from the taxonomy of behavior change.

Automated Analysis:
The F-measures, precision and recall metrics for the cross validation of the machine learning technique J48 tree were 0.77, 0.79 and 0.77 respectively. The F-measures were calculated as an average of the individual binary classifiers using RRI vectors as attributes for the machine learning algorithm. The reliability measure between rater 1 and the automated classification system is 0.71, rater 2 and the system is 0.736. Therefore, the average system-rater agreement 0.72 approached inter-rater agreement of 0.74. A detailed characterization of thematic distribution over the years 2007 and 2014 as obtained using the automated classification system is shown in Figure 3. Although the users under consideration in these two time periods may not be the same, focusing on proportions of messages across themes over two different timelines gave us an overview of temporal patterns underlying thematic content of QuitNet user interactions.

\[
\text{Percentage of messages in a specific theme in a given year} = \frac{\text{(Number of messages in that theme in that year)}}{\text{Total number of messages in that given year}} \times 100
\]
Table 1. Qualitative analysis of QuitNet messages using the taxonomy of behavior change techniques

<table>
<thead>
<tr>
<th>Themes</th>
<th>Sample Message</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goals and planning</strong></td>
<td>Good morning X and YYY.....I pledge not to smoke today and extend my hand to the next quitter who drops by....:///ABC/day 4</td>
</tr>
<tr>
<td><strong>Feedback and monitoring</strong></td>
<td>You'll be fine. It takes some time for your body to heal !! Just hang in there and we will help you the best we can.:///XXX///36 days, 13 hours, 26 minutes and 28 seconds smoke free. 914 cigarettes not smoked. $208.12 and 6 days, 23 hours of my life</td>
</tr>
<tr>
<td><strong>Social Support</strong></td>
<td>I want to pledge again today too! Thanks for the support, I will not be smoking today, and I offer my hand to the next in line.:///I read YYY is also having a big storm with power outages. We may not see ZZZ today - no power, no computer. Everyone</td>
</tr>
<tr>
<td><strong>Shaping Knowledge</strong></td>
<td>Read profiles and journals.//Learn as much as you can about what you are going to face.//Don't walk into this thinking it is going to be easy or cute, this will be one of the toughest fights of and for your life.//Make the quit the most important thing in</td>
</tr>
<tr>
<td><strong>Natural Consequences</strong></td>
<td>good morning, mine is my smell. I can't believe the other day when it was windy early i could smell my own shampoo and conditioner. And also that I am going to have more time on the earth with my family and friends.:///15 days, 10 hours, 3 seconds smoke free</td>
</tr>
<tr>
<td><strong>Comparison of behavior</strong></td>
<td>That would be me. I know I won't be smoking today. Heck no. Here is my hand for the next in line.../../../XXX//Day 672</td>
</tr>
<tr>
<td><strong>Associations</strong></td>
<td>I am a nonsmoker.///I became a nonsmoker the day I quit, Feb 6th 2007.///</td>
</tr>
<tr>
<td><strong>Repetition and substitution</strong></td>
<td>((((((XXX))))))))///Nice warm fire tonite and it's just the ticket for these chilled bones!! I'll be sending 33,657 unsmoked sick sticks to a blazing end. I don't want or need them!!!///I need a shot of apricot brandy and a relaxing hammock to settle</td>
</tr>
<tr>
<td><strong>Comparison of outcomes</strong></td>
<td>Congratulations everyone.///XXX///39 days, 8 hours, 10 minutes and 22 seconds smoke free. 787 cigarettes not smoked. $2,184.00 and 6 days of my life saved! My quit date: 2/6/2007</td>
</tr>
<tr>
<td><strong>Reward and threat</strong></td>
<td>Great job guys! Congratulations!!!///XXX//</td>
</tr>
<tr>
<td><strong>Scheduled Consequences</strong></td>
<td>///HiXXX://Hey I will gladly take those 571 unsmoke cigarettes for the freedom flames :)////Thank you so much for bringing them. Now what can I get you to eat and drink while you relax and enjoy you</td>
</tr>
<tr>
<td><strong>Self-belief</strong></td>
<td>&quot;I'm one puff away from a pack a day&quot; which, I chanted to myself every morning in my quit - and it is so true, don't you think?///I've been tempted to just have one &quot;drag&quot;, or one &quot;little taste&quot;, but, I know, that I can't..... and won't!!//Won't Ever</td>
</tr>
<tr>
<td><strong>Covert learning</strong></td>
<td>Good morning ladies, I hope you are all well.//The smokefree zone is open, drinks &amp; Lemon Syrup Cake are ready for you.///I pledge not to smoke today, offering my hand in friendship and support to the next.///XXX//D833</td>
</tr>
<tr>
<td><strong>Regulation</strong></td>
<td>Go to the gym. Play a sport. Get adreneline rushing through your veins somehow. That's the only way I've found to release the anger and frustration.///and BTW - you're boss probably NEEDS to hear that he's a moron!!!Just as long as you're not smoking</td>
</tr>
<tr>
<td><strong>Antecedents</strong></td>
<td>I've only had about 5 drinks since I quit.///I think if I had more than two drinks at one time I would definitely smoke.///It just goes hand in hand with me. //Maybe in time. I really don't miss drinking that much, and I used to drink quite a bit. I'm</td>
</tr>
<tr>
<td><strong>Identity</strong></td>
<td>I am a nonsmoker.///I became a nonsmoker the day I quit, Feb 6th 2007.///A non smoker is defined as one who does not smoke, and I do not./I know if I ever smoked again I would be a smoker, and that is what I do not want, so I do not smoke.///I smoke in the p</td>
</tr>
</tbody>
</table>

The themes ‘feedback and monitoring’ and ‘comparison of outcomes’ were the most commonly used themes among QuitNet users in the years 2007 and 2014. The theme ‘comparison of outcomes’ has increased in proportion by 22% from 2007 to 2014. This could be attributed to the traditions within QuitNet community such as virtual bonfires where users account for unsmoked cigarettes and discuss consequent benefits with respect to quality of life. The increase in proportion of messages in this theme indicates that the users were aware of the outcomes of the behavior change (both positive and negative), which in turn could have motivated them to remain abstinent. The other theme which has increased in proportions of messages are the ‘reward and threat’, where the increase is 15%. Over the years, QuitNet
users began each of their messages by congratulating the recipient for their quit, which seems to be another new tradition, thus increasing content in this theme. The theme with most dip in percentage of messages were ‘comparison of behavior’ by 9.31%. Although the traditions on QuitNet were the same between the two years, huge reduction in the proportion indicates that number of users participating in the traditions may have reduced. The messages contributing to this theme are those in which users symbolically give their hand to others as a form of support and pledge not to smoke for the day. The messages in the themes ‘feedback and monitoring’, ‘natural consequences’ and ‘goals and planning’ have decreased by around 1%. Messages in which users account for unsmoked cigarettes and discuss the subsequent benefits in terms of quality of life were classified under “self-monitoring”. The proportions of messages in the theme ‘social support’ has reduced by 7.72%. The proportion of messages in this theme was low overall in comparison to other themes across both years. This is not consistent with the results of the previous work done [10] and can be construed as a coding artifact of using the behavior change taxonomy.

Figure 3. Distribution of messages across each theme after automated analysis over the years 2007 and 2014

Overall, this method has enabled visualization of the change in QuitNet users’ communication across time. In summary, machine-learning techniques (to assign categories) in conjunction with distributional models (to provide additional semantic information not available in the content of short social media messages) have facilitated the extension of manually-coded thematic mapping to a large dataset with reasonable accuracy and reliability measures. This task would otherwise be prohibitively labor and resource intensive. Applying distributional semantic techniques such as RRI, to the content on social media has revealed implicit relationships between the messages without dramatically increasing computational overhead.

Temporal Modeling:
For the purpose of temporal modeling across active smoking and abstinence behavioral states, we chose to focus on QuitNet users who relapsed during the study period. Figure 4 portrays the change in percentage of messages before and after relapse.

For a specific theme:

\[
\text{Percentage of messages before relapse} = \left( \frac{\text{Number of messages before relapse in a specific theme in a given year}}{\text{Total number of messages in the theme in that year}} \right) \times 100
\]

\[
\text{Percentage of messages after relapse} = \left( \frac{\text{Number of messages after relapse in a specific theme in a given year}}{\text{Total number of messages in the theme in that year}} \right) \times 100
\]

The percentage difference between message frequency before and after relapse across all themes in the year 2007 was 80%, indicating an overall drop in frequency of communication. This pattern can also be seen in 2014 with a drop of 60% in proportion of messages before relapse and after relapse across all themes. The drop in percentage of messages before and after relapse in the year 2007 was 75% in the categories ‘goals and planning’ and ‘feedback and monitoring’, 66.6% for ‘Social Support’, ‘comparison of behavior’ and ‘comparison of outcomes’, 38.5% for ‘natural consequences’, 71% for ‘miscellaneous’ and 100% for ‘reward and threat’ and ‘self-belief’. The pattern of communication among relapers in the year 2014 was similar to the year 2007. The difference in proportion of
messages before and after relapse in the year 2014 is 70% in the category ‘goals and planning’, 80% in ‘feedback and monitoring’, 70% in ‘social support’, 60% for the categories ‘comparison of behavior’ and ‘comparison of outcomes’, 50% for the theme ‘natural consequences’, 95% for ‘miscellaneous’, 75% in ‘rewards and threat’. Most frequently found themes embedded in communication of relapsing QuitNet users in 2007 belonged to the category ‘feedback and monitoring’ both before and after the relapse event. The theme that had the most significant reduction in number of messages before and after relapse, in the year 2007, was ‘rewards and threat’. Most QuitNet messages which fell under this category are congratulatory in nature for the efforts of the users. Thus, it is coherent that number of messages in this thematic category have reduced. In 2014, the categories most prominent in users’ messages before the event of their relapse, were ‘feedback and monitoring’ and messages that were in the ‘miscellaneous’ category (no relation to smoking cessation in particular). Whereas, the theme that had a huge fall in message count after relapse event in 2014, was ‘miscellaneous’ category. The messages in ‘miscellaneous’ category in QuitNet were related to postings on addictions other than smoking and everyday generic experiences of one’s daily life. Examples of such messages were “my hd has gone out to get a dog”, “Quit drinking without AA! From XXX on 4/9/2014 2:00:48 PM”.

Figure 4. Pattern of communication of relapsing users before and after the event of their relapse

Both in 2007 and 2014, the message count in the thematic category ‘self-belief’ have remained low in comparison to count in other themes, before and after relapse. In 2014, specifically, message count in this theme had remained at 0% before and after relapse. Existing literature of behavior change suggest that users displayed lower self-belief when they were closer to a relapse episode and lower self-belief has been found to be an indicator for relapse [35]. It can also be seen that the number of messages in the category ‘social support’ have also reduced in both years, before and after relapse. The reduction in ‘social support’-related messages after relapse may be attributed to the messages coded under other themes due to explicit discussions of ideas related to quit rebound and coping with failure. Although these messages offered moral support implicitly, they were categorized to ‘shaping knowledge’ and/or ‘natural consequences’ according to the taxonomy of behavior change theories. The overall drop in frequency of communication across both years, has indicated that users tend to communicate less after their relapse. Conversely, when a user quiets down it could indicate that they may relapse. Literature suggests that after relapse, users tend to be affected by guilt and shame, consequently tend to become reclusive [36], which can be observed from the pattern of communication of users over time. Again, the attitude of guilt and shame among relapsers after a relapse event is re-affirmed by the significant drop in messages in the ‘feedback and monitoring’ category. The users may not want others to revisit the event of their relapse unless they cope with their failure. The reduction in message count in the category ‘comparison of behavior’ may also be attributable to the shame that relapers experience after quitting. Overall low presence of the theme ‘goals and planning’ after a relapse event could suggest that the relapers may not be ready to set immediate quit goals [37].

When the post-relapse activity of users was represented as a percentage of the pre-relapse activity, the results revealed interesting insights into specific theoretical techniques embedded in the messages exchanged by the relapers before
and after relapse. In the year 2007, techniques such as ‘goals and planning’ and ‘feedback and monitoring’ were retained in 14.3% messages exchanged post-relapse as compared to pre-relapse. The retention observed was highest in the themes ‘natural consequences’ at 44.4%, followed by ‘comparison of behavior’, ‘comparison of outcomes’ and ‘social support’ at 20%. Themes such as ‘reward and threat’ and ‘self-belief’ had the lowest retention in messages post-relapse. In the year 2014, themes that were retained in the messages post relapse were highest in ‘natural consequences’ at 33.3%, followed by ‘comparison of behavior’ and ‘comparison of outcomes’ at 25%. The messages in ‘Miscellaneous’ category were retained in 2.7% of the messages and was the lowest post-relapse retention in the year 2014.

In summary, understanding semantics and behavior change techniques before and after relapse has provided us with insights into the specific content that users might be interested in and can benefit from, at points in time corresponding to changes in behavioral states within a smoking cessation episode. Such information is essential to design targeted user-content and interactions to enhance the efficacy of existing health promotion platforms such as QuitNet. The information on the users’ low self-efficacy and their attitude towards recovery is evident from analyzing the semantic content. The reduced communication indicating guilt and shame has been observed from the reduced frequency of interaction and also from the semantic content attributes.

**Limitations and Future Work:**

The BCT taxonomy used in this study specifically targets the identification of theoretically driven techniques that target a particular health behavior, in our case abstinence from smoking. However, applying the BCT taxonomy to analyze user-generated content and peer interactions in QuitNet like online platforms may result in omission of important social interactions that, though not recognized as techniques of behavior change, foster trust, bonding, and nurturing of the community. These aspects are ancillary to smoking cessation, yet are important mediators of behavior change in social platforms. Therefore, if it is to be used to annotate such data, the taxonomy should be extended to incorporate these interpersonal interactions – particularly as these may include community-initiated traditions that the users have developed over years of collective effort toward change in behavior. These are best captured through inductive coding techniques such as grounded theory. Our previous and ongoing work employs these techniques to capture emergent nature of user communication within QuitNet [26, 27, 38]. Certain themes did not have adequate representation during qualitative coding, so could not be considered for automated analysis on account of a lack of training cases. Other methods of representation learning (e.g. word embeddings [39]), machine learning involving deep neural networks [40]) may enhance the ability of our methods to generalize from small numbers of training examples. We have used data from two different years 2007 and 2014 to validate the generalizability of thematic categorization, scalability of automated methods, and observed semantic shifts before and after relapse in temporal models. In our future work, we will extend the study to incorporate additional datasets to examine continuous temporal trends. Such analysis will help us to identify semantic content that is predictive of behavior change, and vice versa. Relapse is an important behavioral phenomenon from a clinical perspective. Characterization of the temporal and social dimensions of relapse may permit proactive identification of relapsing users, with subsequent personalization of interventions to meet the associated user needs and behavioral targets. Understanding other user behavioral states beyond relapers (e.g. active smokers, successful quitters) is equally important to help us distinguish specific communication characteristics underlying peer interactions based on smoking status.

**Conclusion:**

This study focuses on using peer-to-peer communication in an online community to understand the implications for data-driven health promotion. The major contributions of this study are as follows:

- This study uses a taxonomy of behavior change techniques to annotate peer communications in a health-related online community for smoking cessation. This offers an insight into the theoretical roots and related techniques embedded in QuitNet user communication.
- The automated analysis methods were used to extend the annotation from a set of 1000 manually coded messages to a large data set of about 84,204 messages. Use of distributional information that captures implicit meanings associated with peer interactions has provided us with better understanding of semantic context within terse social media interactions.
- The temporal modeling of QuitNet peer interactions has focused on understanding the semantic context surrounding a user’s communication across multiple behavioral points surrounding an event of relapse.
semantic context offers insights into users’ needs and triggers, which may be used for effective design of interventions appropriate to a particular stage of behavior change.

Annotating peer-to-peer communication to relate it to established theories of behavior change provides a bridge between these empirical data and scientific understanding of the mediators of behavior change. To our knowledge, this is the first effort to use the taxonomy of behavior change techniques (version 1 with 16 themes) to evaluate the construct of social media as a venue for behavior change. The application of the taxonomy for evaluation has revealed the theoretical manifestations underlying QuitNet user interactions that are primarily user-driven rather than expert-guided. Such findings can help us design better computer-mediated support technologies that nudge and prompt users to exchange theory-guided messages with embedded behavior change techniques. Distributional semantics in conjunction with machine learning, reduces human effort in dealing with large corpus of social media data making it amenable for large scale temporal modeling. Longitudinal trends in communication help us understand the traits of the user groups (classified based on their smoking behavior) that can help us customize the support infrastructure on social platforms for just-in-time guidance to sustain long term behavior change efforts. Indicators of relapse and understanding of user attitudes before and after relapse are essential to provide tailored assistance at the point in time when it is most relevant to the users of QuitNet like behavior change technologies in the digital era.

Acknowledgements:

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number 1R21LM012271-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

3. Zhang M, Social media analytics of smoking cessation intervention: user behavior analysis, classification, and prediction. 2015; Drexel University, Pennsylvania
8. Thrul J., Klein BA, Ramo DE. Smoking cessation intervention on facebook: which content generates the best engagement?. Journal of Medical Internet Research. 2015; 17(11). e244. doi:10.2196/jmir.4575


32. Yuan Lu, Application of random indexing to multi label classification problems: a case study with mesh term assignment and diagnosis code extraction. 2015. University of Kentucky, Kentucky


A Social Network Analysis of Cancer Provider Collaboration

Bryan D. Steitz1, Mia A. Levy, M.D., Ph.D.1,2
Vanderbilt University Medical Center, 1Department of Biomedical Informatics, 2Department of Medicine, Division of Hematology and Oncology, Nashville, TN

Abstract
Cancer treatment often consists of multiple therapeutic modalities delivered by specialists. As changing reimbursement paradigms move towards quality outcomes and bundled payments, extensive care coordination between healthcare providers is imperative. We developed an approach to quantify care coordination relationships among providers treating breast cancer patients at the Vanderbilt University Medical Center. Our cohort of 1285 providers treated 3924 breast cancer patients, and had 1758 unique provider-provider relationships. Providers treating stage III breast cancer patients had the highest ratio of providers to patients, indicating a more tightly connected network than providers treating stage I or II patients. Network analysis can provide quantitative approaches to understanding the relationships of multi-specialty providers and may inform approaches to measuring the impact of care coordination on outcomes.

Introduction
Cancer prevalence is growing, with nearly 1.7 million new cases expected in 2016. Breast cancer is the second most common form of cancer, with nearly 250000 new cases expected in 2016. Cancer management is complex, requiring multiple treatment modalities across diverse settings, managed by many healthcare providers who must coordinate care over time. Care coordination is a multidimensional concept involving the integration of care across all providers and settings. Coordination involves information sharing among the patient, clinicians, and care providers to ensure effective communication of accurate information. A previous study by Smith and colleagues found that cancer patients see an average of 32 physicians over the course of their treatment. Without appropriate coordination among these providers, patients can experience treatment delays, poorer outcomes, and inevitably higher costs.

Cancer care coordination has received attention as an approach to deliver high-value care. With cancer treatment complexity growing, the cost of cancer is projected to reach nearly $158 billion by 2020. A study by Ekwueme and colleagues estimated that many Medicaid breast cancer patients incurred over $5700 in direct monthly costs while receiving their treatment. Similarly, Pollack and colleagues found that patients who received care from a connected network of physicians had lower care costs than patients who visited providers with a less connected network. As reimbursement paradigms shift to a value-based model focusing on quality outcomes and bundled payments, extensive care coordination among specialists is imperative.

Analyzing provider relationships as a social network, or network of interactions between providers, is one methodology used to evaluate coordination and collaboration. In one study, researchers observed a survival advantage in stage III colon cancer patients when medical oncologists and surgical oncologists shared at least three patients. However, to evaluate adequately the significance of tightly coordinated networks of providers on cancer patient outcomes, we must first devise methods to describe and measure the connectedness of provider networks. In this study, we evaluate the network defined by collaborations between providers treating stage I through stage III breast cancer patients. We employ a network analysis methodology to quantify the collaboration between providers treating stage I through stage III breast cancer patients. We define provider collaboration as the number of breast cancer patients shared between two providers.

Methods
This study was conducted at the Vanderbilt-Ingram Cancer Center (VICC) at Vanderbilt University Medical Center (VUMC), an academic health care center in central Tennessee and a major referral center for the Southeastern United States. We collected data on breast cancer patients who met criteria for inclusion in the VUMC tumor registry; those who had been diagnosed or had received all or part of their first course of treatment at VUMC. Data in the VUMC tumor registry follows the North American Association of Central Cancer Registries data standards and dictionary schema. The Vanderbilt University Institutional Review Board approved this study (Protocol 130957).
**Study Population**

Patients with stage I, stage II, or stage III breast cancer diagnosed between January 1, 2000 and December 31, 2014 were included in the cohort. Data on patient demographics, diagnosis characteristics, and treating provider characteristics were extracted from the VUMC Tumor Registry. Demographic data included patient race, sex, ethnicity, and age at diagnosis. Diagnosis-related data included date of initial cancer diagnosis, summary of cancer treatments, date of each treatment, and cancer stage. Provider characteristic data included name, national provider identifier (NPI), and facility associated with each treatment for each patient.

**Provider Specialty Identification**

We used the provider characteristics from our initial data extraction to create a list of all providers and their respective NPI and treatment location (Figure 1). We downloaded the January 2016 NPI registry file\(^{13}\) and imported it into a PostgreSQL\(^{4}\) database. The NPI registry is a national database of medical provider identifiers. We chose to use the NPI registry to obtain provider information to ensure the accuracy of provider specialties, which was not as well represented in the tumor registry data. For each provider in our dataset without a listed NPI, we queried the NPI registry by provider name and location and recorded each match. Queries yielding duplicate name possibilities for the given location were recorded for manual review. Providers without any matches were separately recorded for manual review to account for potential data entry errors.

To validate all pre-populated NPI numbers, we queried the NPI registry. Each NPI number without a successfully matched name was flagged for manual review. For each provider flagged for review, we manually queried the NPI registry for a match. We reviewed each query result for potential matches. Each successfully matched provider was recorded in our dataset. Providers without clear matches, such as individuals with the same name at the same location were indicated as having an unknown NPI number. We excluded providers without an NPI number for subsequent specialty identification.

We queried the NPI registry to extract specialty codes for each provider. In cases where providers had more than one medical license, we extracted codes associated with their current state. The extracted specialty codes were translated into specialty names using the Centers for Medicare and Medicaid Services taxonomy definitions\(^{15}\). We manually reviewed the list of provider specialties to determine larger, more general, specialty categories for network creation. Providers categorized as medical oncologists, radiation oncologists, or surgical oncologists were included for network representation due to their relevance in breast cancer treatment. Provider specialties composing the medical oncology, radiation oncology, and surgical oncology categories are shown in Table 1.

\[
\text{Figure 1: Procedure for determining National Provider Identification (NPI) numbers.}
\]
Table 1: Custom mapping of Centers for Medicare and Medicaid Services taxonomy definitions to respective cancer specialties including Medical Oncology, Radiation Oncology, and Surgical Oncology Categories. Taxonomy codes are represented in parentheses next to each of the specialties.

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Taxonomy Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Medical Oncology</strong></td>
<td></td>
</tr>
<tr>
<td>Internal Medicine, Medical Oncology</td>
<td>(207RX0202X)</td>
</tr>
<tr>
<td>Internal Medicine, Hematology &amp; Oncology</td>
<td>(207RH0003X)</td>
</tr>
<tr>
<td>Internal Medicine, Hematology</td>
<td>(207RH0000X)</td>
</tr>
<tr>
<td><strong>Radiation Oncology</strong></td>
<td></td>
</tr>
<tr>
<td>Radiology, Radiation Oncology</td>
<td>(2085R0001X)</td>
</tr>
<tr>
<td><strong>Surgical Oncology</strong></td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>(208600000X)</td>
</tr>
<tr>
<td>Surgery, Surgical Oncology</td>
<td>(2086X0206X)</td>
</tr>
</tbody>
</table>

Network Representation

We created social networks, or networks of relationships between physicians, to identify collaboration patterns among medical oncologists, radiation oncologists, and surgical oncologists. Each of the social networks consists of nodes, or circles on the graph, and edges, or lines between nodes. Nodes represent providers associated with the care of a patient. Edges represent a relationship between two providers defined by the fact that they share in the care of an individual patient. In typical network diagrams, node circle size and edge line thickness in a network diagram may be modulated to represent the magnitude of what they represent. In this study, the size of each node represents the total number of patients treated by the provider and the thickness of each edge represents the total number of patients shared between two providers. Nodes are color coded by provider specialty to understand inter-specialty relationships.

To create networks, we combined the lists of patients and providers to create a table of unique patient-provider pairs. Providers associated with each patient were combined into provider-provider relationships such that each provider associated with a patient was paired with every other provider associated with that patient. The resulting provider-provider relationships were reduced to the set of unique relationships and a count of the occurrences of that relationship determined the respective thickness or weight of the edge between two providers.

Network Visualization

We created two types of network visualizations: 1) a large, interconnected, network of all medical oncologists, radiation oncologists, and surgical oncologists who treated patients in our cohort, and 2) individual provider networks for top volume providers in each specialty. We used the igraph package within R 3.2.0 to create and visualize the networks. Network layouts were determined using the graphopt, force-directed, algorithm.

Network Analysis

For each network, we calculated the number of patients and providers included in the graph, and the number of relationships between providers. Node and edge sizes were summarized with means, medians, and interquartile ranges. We analyzed the relationship between the number of patients and number of providers in each network to normalize the relative collaboration between providers. We also calculated provider influence within each of the networks by measuring the percentage of patients seen by each of the providers. By evaluating each node’s color within the network, we can identify provider significance within a particular specialty and collaboration between specialties within the context of all providers.

We also created individual provider networks to evaluate collaboration patterns between individual providers. To determine potential referral patterns, we analyzed provider relationships across specialties. We similarly analyzed inter-specialty relationships. Finally, we compared intra-institution and inter-institution collaborations between providers.

Results

Our data included 3924 breast cancer patients with stage I-III disease who received treatment from at least one VUMC-affiliated provider between January 1, 2000 and December 31, 2014, and who had a medical oncologist, radiation oncologist, or surgical oncologist documented in the VUMC tumor registry. Table 2 presents the number of patients who had zero, one, two, three, or more VUMC-affiliated surgical oncologists, medical oncologists, and
radiation oncologists recorded in the tumor registry. On average, patients in the VUMC tumor registry had 2.17 (range 1 to 6) VUMC-affiliated providers listed with a median of 2 providers. Some (7%) patients in the tumor registry were only diagnosed at VUMC and never received treatment there.

**Table 2**: Number of VUMC-Affiliated Providers Treating Each Patient. Percentage representations under each specialty designate the percentage of total patients receiving treatment from a provider of that specialty.

<table>
<thead>
<tr>
<th>Total Number of Patients (%)</th>
<th>Surgical Oncologist (%)</th>
<th>Medical Oncologist (%)</th>
<th>Radiation Oncologist (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 VUMC Affiliated Providers</td>
<td>276 (7.0)</td>
<td>173 (4.4)</td>
<td>171 (4.4)</td>
</tr>
<tr>
<td>1 VUMC Affiliated Provider</td>
<td>706 (18.8)</td>
<td>432 (11.0)</td>
<td>208 (5.3)</td>
</tr>
<tr>
<td>2 VUMC Affiliated Providers</td>
<td>1635 (41.7)</td>
<td>1475 (37.6)</td>
<td>1505 (38.4)</td>
</tr>
<tr>
<td>3 VUMC Affiliated Providers</td>
<td>1092 (27.8)</td>
<td>1157 (29.5)</td>
<td>1170 (29.8)</td>
</tr>
<tr>
<td>&gt; 3 VUMC Affiliated Providers</td>
<td>215 (5.5)</td>
<td>304 (7.8)</td>
<td>325 (8.3)</td>
</tr>
</tbody>
</table>

Figure 2 visualizes the cancer provider collaboration network for surgical oncologists, medical oncologists and radiation oncologists treating the 3924 stage I-III breast cancer patients. The entire network consists of 409 providers with 1758 unique provider-to-provider collaborations. Network statistics for each of the stages are shown in Table 3. More providers (276) treat stage II patients, and have more provider-provider collaborations (885) than either of the other stages. Across each of the stages, medical oncology has the highest number of providers with 166 total medical oncologists. Radiation oncology is the least abundant specialty with 92 total radiation oncologists. The provider network for stage III breast cancer patients has the largest provider-patient ratio with 0.31 patients per provider, whereas the respective networks for stage I and stage II patients have ratios of 0.12 and 0.20. One radiation oncologist dominates the network, treating over half of the patients who receive that treatment (Table 4). The second largest volume radiation oncologist treats only 9.5% of the patients. The top medical oncologist and surgical oncologist treat 21.4% and 21.5% of patients respectively.

**Figure 2**: Cancer Provider Collaboration Network. Each node (circle) represents a unique provider in the network. The color of the node represents the type of cancer specialist: surgical oncology (black), medical oncology (white), and radiation oncology (grey). The size of the node represents the relative number of patients treated by each provider. Edges (lines) between nodes in the network represent provider-provider relationships created when two providers care for the same patient. The weight (thickness) of the edge represents the number of patients shared between two providers. In this network, 409 providers treated 3924 breast cancer patients diagnosed with stage I-III disease between 2000 and 2014. This network contains 1758 unique provider-to-provider relationships with an average of 3.7 (range 1 to 212) patients per provider-provider relationship.
Table 3: Breast cancer provider network statistics by stage

<table>
<thead>
<tr>
<th></th>
<th>Stage I</th>
<th>Stage II</th>
<th>Stage III</th>
<th>Stages I-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients (%)</td>
<td>1985</td>
<td>1399</td>
<td>540</td>
<td>3924</td>
</tr>
<tr>
<td>Number of Providers (%)</td>
<td>242</td>
<td>276</td>
<td>199</td>
<td>409</td>
</tr>
<tr>
<td>Surgical Oncology (%)</td>
<td>90 (37.2)</td>
<td>104 (37.7)</td>
<td>64 (32.2)</td>
<td>151 (36.9)</td>
</tr>
<tr>
<td>Medical Oncology (%)</td>
<td>93 (38.4)</td>
<td>119 (43.1)</td>
<td>87 (43.7)</td>
<td>166 (40.6)</td>
</tr>
<tr>
<td>Radiation Oncology (%)</td>
<td>59 (24.4)</td>
<td>53 (19.2)</td>
<td>48 (24.1)</td>
<td>92 (22.5)</td>
</tr>
</tbody>
</table>

Unique Edges

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (range)</td>
<td>862</td>
<td>885</td>
<td>598</td>
<td>1758</td>
</tr>
</tbody>
</table>

Provider Node Size

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (range)</td>
<td>17.38 (1, 453)</td>
<td>10.9 (1, 287)</td>
<td>6.45 (1, 145)</td>
<td>20.8 (1, 885)</td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Edge Size

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (range)</td>
<td>3.5 (1, 105)</td>
<td>2.55 (1, 83)</td>
<td>1.89 (1, 36)</td>
<td>3.7 (1, 212)</td>
</tr>
<tr>
<td>Median</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Ratio of Providers to Patients

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of Provider Edges to Patients</td>
<td>0.12</td>
<td>0.20</td>
<td>0.37</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Across the entire network, only 6% (25) of the providers are affiliated with VUMC. Provider-provider collaborations including one VUMC-affiliated provider account for 80.5% of the total collaborations in the network. Similarly, 9% of the collaborations are between two VUMC-affiliated providers. The collaborations between two VUMC-affiliated providers account for 55.5% of the total edge weights. Among VUMC-affiliated providers, an average of 74% (range 51.2%, 100.0%) of the weighted edges are with another VUMC-affiliated provider (Table 5). The majority (60.4%) of collaborations between two VUMC-affiliated providers share more than three patients, while only 4.1% of collaborations between non-VUMC-affiliated providers share more than three patients. Similarly, 21% of collaborations between two VUMC-affiliated providers share one patient, while 85.7% of non-VUMC-affiliated providers share one patient.

Figure 3 shows individual provider networks for the highest volume cancer provider in each specialty. Summary statistics are presented in Table 6. The top medical oncologist treating 723 patients has the most collaborators, with 159 unique provider-provider relationships. The top radiation oncologist and surgical oncologist have 140 and 131 provider-provider relationships respectively.

Over 28% of provider-provider relationships for both the top medical oncologist and top surgical oncologist are intra-specialty. 21.3% of the top radiation oncologist’s provider-provider relationships are intra-specialty. The top surgical oncologist shares the most patients within the same specialty, accounting for 8% of the total number of shared patients. The top surgical oncologist and radiation oncologist respectively share 5.5% and 3.8% of patients within the same specialty.

The top radiation oncologist shared four or more patients in one quarter (24.8%) of the provider-provider relationships. Four or more patients were shared in 16.7% of the top radiation oncologist’s intra-specialty relationships. The top surgical oncologist and medical oncologist shared four or more patients in 18.9% and 13.8% of relationships respectively. Four or more patients were shared in 8.1% of the top surgical oncologist’s intra-specialty relationships and 4.4% of the top medical oncologist’s intra-specialty relationships.

Table 4: Percentage of patients treated by top providers within each specialty.

<table>
<thead>
<tr>
<th></th>
<th>Top Provider (%)</th>
<th>Top Two Providers (%)</th>
<th>Top Three Providers (%)</th>
<th>Top Four Providers (%)</th>
<th>Top Five Providers (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgical Oncology</td>
<td>21.5</td>
<td>41.3</td>
<td>55.9</td>
<td>66.4</td>
<td>73.0</td>
</tr>
<tr>
<td>Medical Oncology</td>
<td>21.4</td>
<td>37.6</td>
<td>53.7</td>
<td>60.1</td>
<td>64.8</td>
</tr>
<tr>
<td>Radiation Oncology</td>
<td>55.9</td>
<td>65.4</td>
<td>71.9</td>
<td>75.5</td>
<td>77.3</td>
</tr>
</tbody>
</table>

1991
Table 5: Summary statistics for VUMC-affiliated providers by specialty.

<table>
<thead>
<tr>
<th></th>
<th>VUMC-Affiliated Surgical Oncologists</th>
<th>VUMC-Affiliated Medical Oncologists</th>
<th>VUMC-Affiliated Radiation Oncologist</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Number of Patients</strong></td>
<td>2622</td>
<td>2639</td>
<td>1205</td>
</tr>
<tr>
<td><strong>Unique Edges</strong></td>
<td>551</td>
<td>739</td>
<td>285</td>
</tr>
<tr>
<td>Within Same Specialty (%)</td>
<td>146 (26.5)</td>
<td>204 (27.6)</td>
<td>39 (13.7)</td>
</tr>
<tr>
<td>With Different Specialties (%)</td>
<td>405 (73.5)</td>
<td>535 (72.4)</td>
<td>246 (86.3)</td>
</tr>
<tr>
<td>With VUMC providers (%)</td>
<td>90 (16.3)</td>
<td>116 (15.7)</td>
<td>75 (26.3)</td>
</tr>
<tr>
<td>Sum of Weighted Edges</td>
<td>3447</td>
<td>3877</td>
<td>2219</td>
</tr>
<tr>
<td>Within Same Specialty (%)</td>
<td>209 (6.1)</td>
<td>292 (7.5)</td>
<td>86 (3.9)</td>
</tr>
<tr>
<td>With Different Specialties (%)</td>
<td>3238 (93.9)</td>
<td>3585 (92.5)</td>
<td>2133 (96.1)</td>
</tr>
<tr>
<td>With VUMC providers (%)</td>
<td>2660 (77.2)</td>
<td>2713 (70.0)</td>
<td>1685 (75.9)</td>
</tr>
<tr>
<td><strong>Ratio of Unique Edges to Patients</strong></td>
<td>0.21</td>
<td>0.28</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Figure 3: Individual cancer provider networks for the top volume surgical oncologists, medical oncologist and radiation oncologist.
Table 6: Summary statistics for the top providers of each specialty

<table>
<thead>
<tr>
<th></th>
<th>Top Surgical Oncologist</th>
<th>Top Medical Oncologist</th>
<th>Top Radiation Oncologist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Patients</td>
<td>761</td>
<td>723</td>
<td>885</td>
</tr>
<tr>
<td>Unique Edges</td>
<td>131</td>
<td>159</td>
<td>140</td>
</tr>
<tr>
<td>Within Same Specialty (%)</td>
<td>37 (28.2)</td>
<td>45 (28.3)</td>
<td>30 (21.4)</td>
</tr>
<tr>
<td>With Different Specialties (%)</td>
<td>94 (71.8)</td>
<td>114 (71.7)</td>
<td>110 (78.6)</td>
</tr>
<tr>
<td>With VUMC providers (%)</td>
<td>19 (14.5)</td>
<td>15 (9.4)</td>
<td>22 (15.7)</td>
</tr>
<tr>
<td>Sum of Weighted Edges</td>
<td>746</td>
<td>1143</td>
<td>1617</td>
</tr>
<tr>
<td>Within Same Specialty (%)</td>
<td>60 (8.0)</td>
<td>63 (5.5)</td>
<td>61 (3.8)</td>
</tr>
<tr>
<td>With Different Specialties (%)</td>
<td>686 (92.0 %)</td>
<td>1080 (94.5 %)</td>
<td>1556 (96.2 %)</td>
</tr>
<tr>
<td>With VUMC providers (%)</td>
<td>524 (70.2 %)</td>
<td>887 (77.6 %)</td>
<td>1271 (78.6 %)</td>
</tr>
<tr>
<td>Ratio of Unique Edges to Patients</td>
<td>0.17</td>
<td>0.22</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Discussion

We have developed a methodology to visualize and quantify cancer provider collaboration networks using tumor registry data for breast cancer patients. Using simple network graph statistics, we are able to quantify the degree of connectedness of a group of specialists providing multi-disciplinary therapy for a specific patient population. Multiple studies have employed social network methodologies to quantify collaborative relationships. A previous study by Bridewell et al. demonstrated the effectiveness of social network analysis in quantifying institutional boundaries between neighboring organizations\textsuperscript{19}. Social network analysis of collaborative relationships has also been applied to non-clinical healthcare domains. Studies by Malin, Carly, and Long et al. have each applied network centrality measures to analyze relationships in scientific communities\textsuperscript{20, 21}. Similarly, Hether et al. used social network analysis to evaluate user interactions on prenatal support websites\textsuperscript{22}.

Our methodology offers a scalable approach to analyze cancer provider collaboration networks. The scalable approach is supported by our use of data from the VUMC tumor registry, which stores data within NAACR guidelines. Data format consistency, between cancer types and other external registries, allows us to extend our methodology to evaluate differences in provider collaboration networks across cancer diagnosis. To normalize network statistics across different networks, we used the ratio of providers to patients and the ratio of provider edges to patients across each stage. The ratio of providers to patients allowed us to measure the relative size of each network. The ratio of provider edges to patients allows us to measure the network’s relative density. Both networks normalize for the relative number of patients. Other common network statistics, such as density and connectivity, failed to normalize population differences across networks.

Our scalable approach is not without limitations. Our data was based on the tumor registry’s knowledge of providers involved in the patient’s care, and may not reflect all cancer providers who cared for each patient. The VUMC tumor registry has received awards for their abstraction process. Less complete tumor registries could limit the generalizability of this approach to other institutions. We could also improve our methodology by adding additional data sources. Bridewell et al. refined their network analysis by incorporating treatment and billing data extracted from electronic health records\textsuperscript{19}. Other studies\textsuperscript{3} have used data from the SEER-Medicare database\textsuperscript{23}, which links tumor registry data to Medicare claims data providing a more complete picture of all of the providers treating Medicare patients.

Furthermore, our study is limited by the fact that it provides a static view of an inherently dynamic system of provider-provider relationships that change over time. Our network analysis contains provider-provider relationships spanning 14 years during which time some providers joined and others departed our institution or the geographical region, thus changing the dynamic of their referral patterns. Dynamic network analysis has previously been used in other domains to study changes in social interactions over time\textsuperscript{24, 25}. Future work will incorporate dynamic network analysis techniques to address network temporality.

Our results indicate that the majority (55 percent) of provider-provider collaborations occur between VUMC affiliated providers. We also found that 74 percent of each VUMC affiliated provider’s interactions are with another VUMC provider. Similar to our results, Bridewell et al. observed strong institutional ties between providers at an academic medical center\textsuperscript{19}. Our results also indicated that 68 percent of the provider-provider relationships

1993
share only one patient, while 13 percent share four or more patients. Similarly, providers shared an average of 3.7 patients with a median of one patient. A study of collaboration between surgical oncologists and medical oncologists treating stage III colon cancer patients in the SEER-Medicare database by Hussain et al. found that nearly three quarters of providers share at least two patients, with a median of three shared patients. However, among relationships between two VUMC providers, we found that nearly 79 percent of the provider-provider collaborations share at least two patients, with 60 percent of relationships sharing more than three patients. These results indicate strong collaboration within VUMC, with many individual collaborations between external providers, which may be related to geographically distanced patients visiting VUMC to receive part of their care, but receiving much of their daily care closer to home. The study by Hussian et al. also correlated a survival advantage with surgical oncologists and medical oncologists who share more patients. Our future work will focus on evaluating the correlation between the number of patients shared between providers and clinical outcomes including patient survival and other process outcomes.

To our knowledge, this study is one of the first to evaluate individual specialist networks. To normalize each individual provider networks by the relative number of patients, we measured the ratio of unique edges to patients for each provider. Our results indicate that the top radiation oncologist has the lowest ratio with 0.16 unique edges per patient, while the top medical oncologist has the highest ratio with 0.22 unique edges per patient. The low ratio for the top radiation oncologist indicates a high level of collaboration between fewer providers; potentially due sharing many patients with other VUMC affiliated specialists. We hypothesize the medical oncologist’s larger ratio is due to geographically distanced patients receiving much of their day-to-day care from providers closer to home.

Within the diagnosis of breast cancer, we were able to identify differences in provider collaboration networks for a sub-population of stage III breast cancer patients. Stage III breast cancer has a higher risk of recurrence and is more often treated with pre-surgical or post-surgical adjuvant chemotherapy than stage I or II breast cancer. The intensity of treatment and coordination with surgical plan management requires a closer collaboration between surgical and medical oncologists. We hypothesized that stage III cancer provider collaboration would be more closely connected than stage I or II provider networks. We observed that stage III breast cancer patients had the highest provider-patient ratio and provider-edge-patient ratio compared to stage I and II. These higher ratios indicate that providers are sharing more patients and are more closely connected than other stages, confirming our hypothesis.

**Conclusion**

Cancer treatment often consists of multiple treatment modalities, managed by many care providers. While improved care coordination has been identified as a way to save costs and deliver high value care, few methods exist to quantify the relationships between multi-specialty providers. We employed a network analysis approach to evaluate the collaborations between surgical oncologists, medical oncologists, and radiation oncologists treating stage I – III breast cancer patients. Not surprisingly, we found that intra-institutional relationships were stronger than inter-institutional relationships. We also found that as cancer stage increases, the ratio of providers to patients increases to better coordinate more complex care. Network analysis can provide quantitative approaches to understanding the provider relationships between specialties and may inform approaches to better understand the impacts of care coordination on patient care.

**Acknowledgements**

We thank Dr. Jeremy Warner and the Vanderbilt tumor registry for their assistance with data extraction. Bryan Steitz was supported by the 5T15LM007450 training grant from the National Library of Medicine.
References

17. RCT. R: A language and environment for statistical computing. [Internet]. Vienna, Austria. Available from: https://www.R-project.org/
18. Schmuhl M. Graphopt [Internet]. Available from: http://www.schmuhl.org/graphopt
Baby Steps Text: Feasibility Study of an SMS-Based Tool for Tracking Children’s Developmental Progress

Hyewon Suh, MS, John R. Porter, MS, Robert Racadio, MS,
Yi-Chen Sung, MS, Julie A. Kientz, PhD

1University of Washington, Seattle, WA

Abstract

To help reach populations of children without consistent Internet access or medical care, we designed and implemented Baby Steps Text, an automated text message-based screening tool. We conducted preliminary user research via storyboarding and prototyping with target populations and then developed a fully functional system. In a one-month deployment study, we evaluated the feasibility of Baby Steps Text with fourteen families. During a one-month study, 13 out of 14 participants were able to learn and use the response structure (yielding 2.88% error rate) and complete a child development screener entirely via text messages. All post-study survey respondents agreed Baby Steps Text was understandable and easy to use, which was also confirmed through post-study interviews. Some survey respondents expressed liking Baby Steps Text because it was easy, quick, convenient to use, and delivered helpful, timely information. Our initial deployment study shows text messaging is a feasible tool for supporting parents in tracking and monitoring their child’s development.

1. Introduction

Approximately 1 in 6 children between the ages of 3 and 18 in the United States are diagnosed with developmental disorders, such as autism, ADHD, or anxiety disorders. It is known that the earlier these disorders can be diagnosed, the sooner early interventions can start, resulting in better outcomes for these children. Unfortunately, approximately half of children who eventually receive diagnoses of developmental disorders are not identified until they reach kindergarten in the United States, with lower-income populations being missed at a higher rate. Tracking and regular screening of child development is the key to early detection of many types of developmental disabilities. Developmental screening involves regular checks for developmental milestones at up to 22 different intervals across a child’s first 5 years of life. Milestone screening includes age-based questions such as “does your baby turn her head towards a loud noise?” and “can your child climb stairs?” Because these activities take place in daily life outside of traditional clinical settings, parents’ and caregivers’ reports for the early signs of developmental delay are the most reliable way to track them. Therefore, it is important to engage, inform, educate, and empower parents to track and assess their child’s development in an easy way in order to make monitoring more accurate and meaningful. Unfortunately, 5 years is a long time to keep parents engaged consistently. In addition, parents without access to developmental screening often have difficulties in monitoring because they may not be aware of what to watch for. Pediatricians provide parents milestone questionnaires to complete at Well Child Visits (regularly scheduled checkups in the U.S). However, given Well Child Visits are months to up to a year apart depending on the child’s age, it is not the timeliest solution. In addition, many resource-constrained populations, such as low-income or immigrant families, lack the insurance to cover Well Child Visits, and frequent moves or language barriers make it difficult to access information about normal child development.

The ubiquity and penetration of mobile phones presents an opportunity, particularly in under-resourced health ecosystems, via SMS (Short Message Service), a low-cost communication channel that is supported on even the most basic of mobile phones. Between 2000 and 2012, the number of mobile phones in use worldwide grew from fewer than 1 billion to around 6 billion. In the United States, 91% of adults own a cell phone as of 2013. This widespread distribution of mobile phones across diverse populations may solve one of the most difficult problems in public health domain – barriers to access. Thus, mHealth (mobile health) can make it possible to inform caregivers and let them collect and share relevant data at any time, allowing more rapid convergence to optimal treatment.

To attempt to reach more underserved populations with child developmental screening, we have developed Baby Steps Text, a tool that prompts parents to track and review developmental milestone data and connect them to resources using only text messages. Baby Steps Text is a completely automated, personalized, and self-contained two-way communication system that also works seamlessly with a larger set of tools that have been developed as part of a developmental screening ecosystem (including a web portal and Twitter-based tools). Baby Steps Text provides a way to understand how we can include resource-constrained populations who may not yet have constant Internet access or sophisticated smartphones. This paper presents the design, development, and feasibility study of the Baby Steps Text system. Our research brings SMS-based mHealth interventions to a new domain (childhood developmental
screening) and helps expand knowledge about how parent-generated health data can be collected in a simple but thoughtful way, thus expanding the reach of its benefits.

2. Related work

2.1. Domains of SMS-based mHealth

Using mobile phones to support health interventions has a number of advantages over other technologies. The biggest advantages of using mobile phones for supporting public health are that they are personal, connected, relatively inexpensive, and always with people. Current evidence suggests that SMS-based mHealth supports the following four health domains: 1) **Disease Management & Treatment Adherence**: Use of text messaging to reinforce treatment (e.g. HIV/AIDS\(^6\), \(^{40}\), diabetes\(^{14},^{24}\), asthma\(^{38},^{42}\), smoking cessation treatment\(^{15},^{37}\), weight loss\(^{34}\), mental illness\(^9\), and medication adherence\(^{29}\); 2) **Health Information Delivery and Behavior Change**: Text messaging as an inexpensive tool for behavior change and the delivery of health information (e.g. educating about pregnancy and newborns\(^{10},^{31}\), vaccination\(^{41}\), youth sexual health\(^{29}\), weight management\(^{16}\), and encouraging physical activity\(^{12}\)); 3) **Patient-Provider Communication**: Use of mobile phone as a communication channel between patients and providers (e.g. clinic appointment reminders\(^5,^{27}\) and communication with visually impaired patients about health information\(^{22}\)); 4) **Early Detection & Screening**: Screening for early detection of disease (e.g. cervical cancer screening\(^{26}\), and breast self exams\(^{26}\)). However, despite advantages of SMS in delivering information at the appropriate time, use of SMS for screening purpose has not been as well-explored as much as other areas. Baby Steps Text fits into this domain most closely, and addresses a new area within it for children’s developmental screening.

2.2. One-way vs. Two-way SMS Communication

One-way SMS communication is similar to mass media in that it “pushes” information to subscribers’ phones by using messages tailored to personal needs. A large body of SMS-based research has typically been designed this way. This includes simple reminder programs such as reminding clinic appointments and other systems such as sending health promotion and dietary and disease management information. In one-way SMS systems, the subscriber is more of a passive recipient rather than an active user. In order to solve this problem, some projects used two-way communication in their communication message design. For users, the interactive nature of two-way communication may require greater effort and generate greater interest\(^{19}\). While some projects support a fully natural way of communication by having a humans responding to questions and concerns sent by users\(^2\), other projects support fully automated communication by requiring users to follow a specific structure to enable automated parsing by the system\(^{1,^{32}}\). Recently, Perrier et al.\(^{55}\) deployed a hybrid two-way communication where the system sends out bulk messages to its users automatically, and a human behind the scenes takes care of user replies.

2.3. Technology Design for Infant and Children’s Health

Researchers in the human-computer interaction community have sought to design technologies in support of children’s health. These areas have primarily been focused on general studies on the opportunities for design of technologies for children’s health, technologies for use of parents and doctors, and technology to be used by children themselves. For general studies of opportunities, Kientz et al. conducted a formative study on the opportunities for the design of developmental tracking systems\(^{21}\), and then later conducted an in-depth evaluation of the opportunities for supporting parent-pediatrician interaction\(^{33}\). Jeong et al. studied parents with the purpose of providing insights for the design of technology for tracking general health data about children\(^8\).

This prior formative work has informed a number of systems designed to support parents in tracking their children’s health. The original Baby Steps system\(^2\) was a software application to track developmental progress, with follow on work expanding it to using Twitter\(^39\). Baby Steps Text seeks to build upon this previous literature by exploring a new delivery mechanism for developmental screening. Additional designs for helping parents monitor their children’s health include ENSURE\(^2\), which help parents track daily health data such as meal intake and diapering, and Fitbaby\(^17\), which helps parents of premature infants track their child’s physical exercises and overall wellbeing. Both ENSURE and Fitbaby focus on general health observations and are not intended for screening, as is the goal of Baby Steps Text.

3. User-Centered Design Process

One of the primary goals for the overall Baby Steps project was to identify ways to reach underserved families for developmental screening. We conducted a user-centered design approach that formed the basis for the final design of Baby Steps Text, which we describe in the following section (see Figure 1) to engage with key members of this initiative to explore the use of technology in helping to reach underserved populations in collaboration with Washington State’s Department of Health. We focused specifically on Latino populations in our state, as they represent a large percentage of underserved populations and experience a number of challenges that may affect other underserved populations, such as access to health insurance, access to mobile technology, and cultural and language barriers\(^3\). However, the intended users of Baby Steps Text are broader than just this specific population.
3.1. User Research

To get a broader perspective on our target population, we conducted semi-structured interviews with five stakeholders, including one Latino parent of a child with a developmental disability, two health care providers, and two Latino community health advocates. The interviews aimed to understand how parents engage in healthcare for their child and participate in developmental screenings and how health information is communicated. The research team members also visited a Latina health fair located in a predominantly working class Latino neighborhood in Seattle. The health fair was organized by a Latino-focused community health center, attended by Latina women, their families, and community leaders. We administered short surveys to 15 mothers and asked questions related to technology use, the number of children in their family, and their language use.

From the interviews, we learned that Latino households are commonly multi-generational—consisting of children, their parents, and their grandparents—and the responsibility of caring for children is shared with members of the extended community. When asked to describe the technologies that members of the Latino community would most likely have access to, the most common answer was the mobile phone with basic capabilities. They suggested that home computers or home Internet access would not be common, but families might use computers in public venues, such as at the library. Findings from the survey supported several findings from the interviews. In particular, mobile phone use was prevalent (93% of respondents), as was text messaging (87%). Only 6 participants (40%) used their phone to access the Internet and only two (13%) downloaded applications, which meant an entirely text message-based system would have broader reach than a mobile website or app, as home computer and Internet access were not as prevalent (60% and 53% respectively). Two mothers (13%) also mentioned they did not use services like mobile Internet or Facebook, but that their spouses did. We learned from the providers and health advocates that they serve a community that varies widely on literacy skills. The Latina fair survey confirmed these numbers, with the majority of participants preferring Spanish, but nearly half also having English proficiency.

To summarize our preliminary user research, we developed a set of 9 personas representing stakeholders of the potential system, including three different family caregiver types, community members, healthcare providers, and advocates in the state (see Figure 2 for an example persona). These personas were used during the ideation and design process to ensure the needs of stakeholders were represented in the design of the tools.

Figure 2. Left: Persona for a parent of a child with special needs. Right: Spanish-language Storyboard of Baby Steps

3.2. Ideation & Prototyping

The research team conducted a design ideation brainstorming session and generated 63 unique ideas for potential technologies to reach Latino families in the state with developmental screening. We kept copies of the personas present during the brainstorming to ensure designs were grounded in the needs of the stakeholders. From there, we narrowed design ideas into the most promising and feasible three ideas for which we developed storyboards in both English and Spanish suitable for user feedback (Figure 2). The three ideas included text messaging, a website, and a public kiosk. We also developed high-fidelity interactive prototypes of a public kiosk and a website using the Axure prototyping software.
tool (http://axure.com) and a Wizard-of-Oz prototype of the text messaging tool using Google Voice to send messages to participants’ phones and a detailed script of messages.

3.3. Evaluation
We reviewed the storyboards and prototypes with stakeholders, including advocates and families attending the Latina health fair. When reviewing the storyboards with potential users, the text messaging-based system seemed to be the most promising approach for the furthest reach. Stakeholders and parents all responded positively to a text-messaging based system, as long as the costs were kept low. Participants were unsure if they would use or have consistent access to a public kiosk, and though a website would be helpful and provide a richer experience, the lack of a home computer or data plans on their phones limited potential access. After deciding to move forward with the text messaging design, we iterated on the design and wording of the Baby Steps Text system through initial Wizard-of-Oz usability studies with 10 parent participants. We had two researchers simulate the experience of using Baby Steps Text for the participant. Participants used their own mobile phones to send and receive text messages to the researcher playing the part of the “wizard,” who used the Google Voice interface on a laptop, along with a script (see Figure 3), to quickly customize responses for the users during testing. We had a total of 12 tasks that covered all aspects of the system, including registering, responding to several milestone questions, receiving a report, requesting help, and stopping the service. Participants had a printed sheet with each task, and a study coordinator guided them through the process while the “wizard” read and responded to the system. Participants offered suggestions on improving the content and process for receiving and responding to text messages. In particular, our original design for the progress report only included the number of questions answered and no real information about how the child was progressing in each category. We thus added more information about the child’s progress and added information about what to do if the child’s result in any given category of development was “More evaluation needed.” The registration component was also confusing, as we required the child’s name, birthdate, and gender to be separated by spaces. We found participants could easily remember the “y” “s” and “n” responses after a few milestone prompts and thus only needed simple instructions rather than the full instructions with each message.

Task 3: Responding to Milestone Questions
Participant task: You will soon receive a message with a question about your child. When you receive the message, respond to it based on your child’s experience using the instructions.
Wizard response: Send message: “Can NAME drink water, juice, or formula from a cup while you hold it? (Reply Y for “yes,” S for “sometimes,” N for “not yet”)”
If response is anything but Y, S, or N, reply back with:
○ Reply to ?’s with “Y” for yes, “S” for sometimes, or “N” for not yet. New msg sent after you reply. Txt “stop” to cancel. Txt “report” for NAME’s progress.

Task 11: Requesting a Report
Participant task: Request a progress report on how many milestones your child has completed so far.
Wizard response: Expect to receive the word “report” from parent. Send back:
NAME’s Progress, Number of milestones completed:
Personal-Social – 34
Gross Motor – 23
Fine Motor – 22
Communication – 40
Problem Solving – 22

Figure 3. Example script for “Wizard” in Wizard-of-Oz prototype for two of the twelve tasks. Wizard instructions are in italics and BOLD CAPS indicate parts customized by the Wizard for the user’s child.

4. Baby Steps Text Design & Implementation
Based on the results of our initial studies, we finalized a fully functional version of Baby Steps Text that would be suitable for a deployment study. The milestone question content used in Baby Steps was inspired by the Ages and Stages Questionnaire. In this section, we describe the design features of the messaging system.

4.1. Personalized Milestone Questions
Several previous studies also have shown the benefits of individually tailored messages that are personalized for the user. Because each message is sent to an individual, we can personalize the content of the messages. Using information from enrollment such as child birthdate, name, and gender, Baby Steps Text can send age appropriate messages and can include the child’s name and gender-specific pronouns (unlike paper-based screeners), conveying a feeling of personal attention cost-effectively. To automate this process, we used placeholders such as %n% and %s% for words that needed tailoring and replaced those via an automated script. Overall, Baby Steps Text consists of 660 distinct milestone questions that are tailored for each child and delivered based on age. Our system differs from many other SMS-based approaches in that all milestone questions are personalized to increase user engagement and is entirely automated to run without behind-the-scenes human effort.

4.2. Simplified Response Structure
There are four response options to milestone questions: ‘Yes,’ ‘Sometimes,’ ‘Not Yet,’ and ‘Revisit’ (all case-insensitive) to skip to the next question and respond to it later. To reduce typing, parents could also use first letter of each response type (e.g., Y for ‘Yes’ and R for ‘Revisit’) (see Figure 4). If the caregiver has multiple children registered with Baby Steps Text, they need to also include their child’s name. Beyond these required elements, caregivers can add free text in their response (e.g., “Yes, Tommy can drink milk from a cup!”) like a natural text conversation, which does not interfere with the system’s automated parsing mechanism. We wanted to repeat instructions for using “Y” “S” and “N” with each milestone prompt to remind participants, but even after shortening, some questions were too long to include the full instructions within the 160-character limit. Thus, for shorter milestone questions, we included the full instructions (“reply ‘Y’ for “Yes”, ‘S’ for “Sometimes”, or “N” for “Not Yet”) whereas for longer questions, we used a simplified instruction (reply ‘Y’, ‘S’, or ‘N’). When we shortened the screener content, we made the longest message a maximum of 130 characters to ensure we had at least 30 characters remaining for the simplest instructions.

4.3. Progress Report
If parents text “Report” to Baby Steps, we generate a child’s progress report based on parents’ responses to individual questions (e.g., whether they are developing on schedule or if they should get in touch with their provider for additional assessment). To fit the progress report into a single text message, we used three simple reporting levels (“On track,” “Let’s watch,” and “More evaluation needed”) for each development category (e.g., Communication, Fine Motor, Gross Motor, Problem Solving, Personal & Social). We aimed to carefully convey progress to reduce unnecessary caregiver anxiety by working on the exact wording with experts a local non-profit organization that provides a free statewide program for family health that serves many low-income families and also runs a toll free parent help hotline to connect parents to early intervention services. If a parent requests a progress report before they answer half of the questions, we provide a message to answer more questions first. If they answer more than half but not all 30 questions, we send an additional message that the progress report they receive may not be an accurate reflection of their child’s development until all questions are completed. If a child is not ‘On Track’ in any development categories based on

![Figure 4. Example of Baby Steps Text syntax usage. Gray (left) indicates messages sent by Baby Steps and green (right) indicates messages sent by caregiver.](image-url)
parents’ responses, we provide the toll free hotline number for the non-profit mentioned above in a message following the progress report so that parents have contact points for questions or concerns about child.

4.4. Intervals of Screening Questions
Because our user research indicated that mobile phone plans used by lower income families may vary and be limited (e.g., pay-per-use plans), we had to find ways to deliver complete screening questionnaires within a certain timeframe so that those screening questions are still valid to the child’s age. We originally intended to send one screening question per day to avoid information overload, but due to this finding, we designed Baby Steps Text to send the next question immediately after they respond to one because it signals that 1) they are active on mobile texting, and 2) they are able to send and receive messages at the current moment. If the parent does not respond to a question within 48 hours, we resend the same question again.

4.5. Supporting Multiple Caregiver Engagement
Child development happens in daily life, and it is not always one parent who observes when their child hits certain milestones. In our user research, we found that often family members shared child-raising responsibility with aunts, uncles, and grandparents, with not every member of the family having consistent access to mobile phones. To coordinate tracking child development across multiple caregivers, we allowed our system to register multiple users’ phones. The same question is sent to all caregivers’ mobile phones, and if one answers a question, the others receive a message saying “[name of other caregiver] answered [response type]” (e.g., “Juan answered yes”) and all will receive the next milestone question to answer. This allows all caregivers follow their child’s development and coordinate responses.

4.6. System Implementation
Baby Steps Text is a custom-built Python application and associated MySQL database, which uses the Twilio cloud communications platform (https://twilio.com) for SMS and MMS handling. Incoming messages sent by users to our phone number are received by Twilio, which then relays the data to our system via HTTP request to be parsed for keywords. Similarly, by sending HTTP requests to Twilio, we can send both SMS and MMS messages to our users programmatically. When our system receives an incoming message, the content is parsed to determine if it is a milestone response or another of our recognized keywords (REPORT, START, STOP, HELP); responses are stored in the relevant child's database entry and sends an appropriate error message to any message that does not fit into the response structure. Outgoing messages are sent in two distinct scenarios. When a user answers a milestone question, the next question is sent immediately. They can also be sent via a daily check (handled by a cron job on the server), which determines on a child-by-child basis if it is time to either send the first question of a new milestone set (if the child has entered a new age range) or resend the last outgoing message (if 48 hours have elapsed without a response).

5. Feasibility Study
In Winter 2015, we conducted a month-long deployment study with 14 participants to test the feasibility and usability of Baby Steps Text. This study included collecting usage logs and qualitative data from pre- and post-study surveys. The study was reviewed and approved as minimal risk research by our university’s human subjects board.

5.1. Participants
Although we did not exclude participants based on income, we aimed to recruit resource-constrained families in our study through the waiting room of local pediatric clinic that serves a high number of underserved, low-income, and immigrant families. Because our current prototype only has English-based questions, we required participants to have basic English literacy. The clinic receptionist introduced eligible parents of children under five years old to our study at check-in and directed them to the researcher in the waiting room to enroll in the study. We recruited 14 participants from a variety of backgrounds and none of feasibility study participants participated in any of our preliminary user studies. The participants were caregivers (M:2, F:12) of at least one child under 5 (Table 1). We aimed to recruit a balance of genders, but in all but three cases (two fathers and one aunt), it was the mother who checked in at the clinic. The majority were aged between 33 and 40 years old (n=8), and the total children in the household varied from zero (aunt) to seven. All owned smartphones and were active text users. Half were single and half were married. Child participant ages ranged from 4 months to 60 months. One participant (P11) used Baby Steps Text with two children.

5.2. Study Procedure
Participants received instructions, signed a consent form, and completed a pre-study survey on family demographics, current mobile phone use, and parenting experience. We also provided a one-page instruction sheet that explained how to use the system. We then registered them with Baby Steps Text, and participants started receiving screening milestone questions. After the first visit, the research team did not interact with participants until one month had passed. After one month, we mailed an exit package that included a study completion brief, compensation, and a post-study survey with a self-addressed, pre-stamped return envelop. Participants received $20 for their study participation.
6. Results

6.1. Baby Steps Text Usage

Among the 14 parent participants, one (P2) withdrew from the study in the early stage by replying “STOP.” The remaining 13 participants successfully used text messages to track their child’s development and completed the study. Over their one-month-long participation, Baby Steps sent 557 messages and received 520 SMS messages in response from participants. On average, each participant who completed the study received 42.3 messages (SD = 18.6) and sent 39.4 messages in response (SD=19.7). The high standard deviation is due to three participants whose child transitioned from one milestone question set age range to another during the course of study (therefore answering 60 questions) and among them, one participant used Baby Steps Text for two children and responded to 90 questions. Four participants (P1, P5, P6, P8) responded to all 30 milestone questions almost in one sitting (Table 2). We did not see a completion rate difference between the two strategies (e.g., responding to all at once vs. periodically).

6.2. Use of Baby Steps Text Syntax

For Baby Steps Text to recognize and store user responses in the database automatically, we required participants to use one of the four response options. We provided an instruction sheet when participants enrolled in our study and continuously included these instructions within the question message itself. As a result, 505 of the 520 user messages were correctly formatted, yielding a very low error rate of 2.88%, and participants were able to self-correct errors early due to an automated response that informed them of syntax errors. About 15% of the text message responses used shortened response types such as Y/S/N, but there was no particular pattern as many participants used both full and

and $10 each for the pre- and post-study survey completion (post-study survey had to be mailed back). Five participants (two were survey respondents and three were not) also participated in the post-study interview to provide us richer feedback. As a token of appreciation, participants received $20 at their sign up to cover any texting fees incurred using Baby Steps Text.

Table 1: Participant details (*A: Asian, B: Black or African American, H: Hispanic or Latino, W: White)

<table>
<thead>
<tr>
<th>ID</th>
<th>Relation to child</th>
<th>Age</th>
<th>Race</th>
<th># of child</th>
<th>Marital status</th>
<th>Education level</th>
<th>Household Income</th>
<th># of texts per day</th>
<th>Child age (mos.)</th>
<th>Child sex</th>
<th>post-study survey</th>
<th>Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Father</td>
<td>33-40</td>
<td>A</td>
<td>1</td>
<td>Married</td>
<td>Some college</td>
<td>$35,000-$49,999</td>
<td>1-10</td>
<td>20</td>
<td>F</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>Mother</td>
<td>33-40</td>
<td>H</td>
<td>7</td>
<td>Single</td>
<td>8” grade or less</td>
<td>&lt; $15,000</td>
<td>1-10</td>
<td>30</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>Mother</td>
<td>33-40</td>
<td>W</td>
<td>3</td>
<td>Married</td>
<td>College degree</td>
<td>&gt; $100,000</td>
<td>1-10</td>
<td>36</td>
<td>F</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>Father</td>
<td>33-40</td>
<td>B</td>
<td>3</td>
<td>Married</td>
<td>Graduate degree</td>
<td>$50,000-$74,999</td>
<td>1-10</td>
<td>20</td>
<td>F</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P5</td>
<td>Mother</td>
<td>25-32</td>
<td>B</td>
<td>-</td>
<td>Single</td>
<td>High school or GED</td>
<td>Under $15,000</td>
<td>1-10</td>
<td>54</td>
<td>M</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>Mother</td>
<td>18-24</td>
<td>B</td>
<td>1</td>
<td>Single</td>
<td>High school or GED</td>
<td>$15,000-$24,999</td>
<td>101-500</td>
<td>27</td>
<td>F</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P7</td>
<td>Mother</td>
<td>33-40</td>
<td>B</td>
<td>6</td>
<td>Married</td>
<td>High school or GED</td>
<td>&lt; $15,000</td>
<td>1-10</td>
<td>14</td>
<td>M</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>P8</td>
<td>Aunt</td>
<td>33-40</td>
<td>A</td>
<td>0</td>
<td>Single</td>
<td>College degree</td>
<td>$50,000-$74,999</td>
<td>11-30</td>
<td>20</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P9</td>
<td>Mother</td>
<td>33-40</td>
<td>W</td>
<td>1</td>
<td>Married</td>
<td>Graduate degree</td>
<td>$75,000-$99,999</td>
<td>1-10</td>
<td>4</td>
<td>F</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>P10</td>
<td>Mother</td>
<td>25-32</td>
<td>H</td>
<td>2</td>
<td>Married</td>
<td>High school or GED</td>
<td>$25,000-$34,999</td>
<td>101-500</td>
<td>60</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P11</td>
<td>Mother</td>
<td>25-32</td>
<td>H</td>
<td>3</td>
<td>Married</td>
<td>High school or GED</td>
<td>$25,000-$34,999</td>
<td>1-10</td>
<td>30</td>
<td>4, F</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P12</td>
<td>Mother</td>
<td>18-24</td>
<td>B</td>
<td>1</td>
<td>Single</td>
<td>High school or GED</td>
<td>&lt; $15,000</td>
<td>1-10</td>
<td>27</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P13</td>
<td>Mother</td>
<td>41-50</td>
<td>B</td>
<td>3</td>
<td>Single</td>
<td>High school or GED</td>
<td>&lt; $15,000</td>
<td>1-10</td>
<td>48</td>
<td>M</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>P14</td>
<td>Mother</td>
<td>33-40</td>
<td>-</td>
<td>3</td>
<td>Single</td>
<td>College degree</td>
<td>Prefer not to say</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Overall text activity by 14 participants. The rightmost cell shows a sparkline plot of the month-long timeline of texting activity. (*P2 withdrew from the study.)
shortened responses interchangeably. Although we included the “Revisit” feature on the instruction sheet, it was not included as a response type in question messages, and thus no participants used this feature during the study. One post-study survey respondent mentioned she did not know about the option but would appreciate having it.

6.3. Participants’ Experience with Baby Steps Text
Six out of the thirteen participants completed and mailed back the post-study paper survey. Meanwhile, five participants participated in a phone interview to share their experience (see Table 1). Although the survey response rate was low, this is not unexpected for mail-in surveys with lower-income populations. All survey and interview participants agreed that milestone questions asked by Baby Steps Text were understandable and the system was easy to use, respectively scoring 4 and 4.3 where 5 is strongly agree. They also agreed the total number of milestone questions (30 per 2-6 months age range) asked was appropriate, which we could also confirmed from interviews. However, P9 was not as satisfied with receiving one message immediately after the other: “the constant stream of texts on the 1st day was a little overwhelming.” This may be due to not clearly explaining they could take time between questions and they would receive reminder messages after 48 hours, which some participants appreciated – “At times I forget to answer and it would hit me up again to make sure I answer them. (P4)” When asked about message frequency, survey respondents were mixed. Half (n=3) preferred to receive the next question immediately and answer in one sitting, and the other half preferred to receive questions periodically. However, most of the interview participants (4 out of 5) preferred to receive questions one after another, as P6 said, “I think that was really cool that you can keep going and don’t need to wait a while for the next question” and P4 also noted, “even though it’s first annoying, I think you should keep them one after the other. My schedule is hectic and if it’s periodic, I am not going to think about it as much.” In their interviews, P4 and P7 mentioned that although milestone pictures were helpful, they had to make space for them because their phone’s storage was full. P7 also noted shortcomings of milestone questions – “There are some didn’t apply because – the ones with stairs, we don’t have stairs. So I couldn’t really answer. Some of them didn’t apply. So I answered ‘no’” For the open-ended survey question about what they liked, P3 and P9 reported that it was easy, quick, and convenient. Several also liked receiving useful information that can be helpful to their child in timely manner (P5, P9, P10). P10 noted, “I really enjoyed participating in this study. I think the texting format is really great and unique. It’s nice to receive personalized texts and not have to go to the Internet to look things up!”

6.4. Continued Usage After Study Completion
The exit package explained to participants that they could choose to continue using Baby Steps Text beyond the study. P10 finished the 60-month milestone questionnaire during our study, which was the last questionnaire of Baby Steps Text system, thus there was no more milestone questions to send. None of the remaining twelve participants who completed the study explicitly opted out of Baby Steps Text. Nine out of 12 participants were still responding to milestone prompts three months study completion, where we received an additional 385 messages from participants. Due to a software glitch, we were not able to send messages longer than 3 months. However, this is promising in that participants found our system engaging and worth using outside of study compensation.

7. Discussion
7.1. Recommended Changes to Baby Steps Text Design
1) More Detailed Progress Report: To fit the progress report into a single text message, we had to make it brief. Although the report design worked well in usability testing, when we provided real data to participants, they were not as satisfied with minimal data. P3 noted that, “I felt the report was too high level and vague. Didn’t tell me much of anything.” To reflect this feedback, the next version of Baby Steps Text will convey more information on interpreting each item by allowing parent to text “more” after a report to receive explanations about each developmental category, what the child’s result means, and what actions they can do beyond calling the family health hotline. For users with Internet access, providing a link to a web-based report with more detail could be useful, as others have done.

2) Providing Ideas for Activities to Encourage Development: To support parents in encouraging development beyond tracking, we plan to implement an opt-in feature where the system sends a message asking if they want to receive ideas for activities they can do to encourage development with their child. If yes, the system would send age-appropriate items automatically (e.g., “Activity Idea: Put dry rice inside an empty bottle to encourage shaking and turning behaviors”). Because activities are suggestions rather than screening, no response would be required.

3) Sentimental Records along with Developmental Records: Previous research indicates that tracking sentimental childhood records, such as memories and “firsts and favorites,” along with development can increase parent engagement. We plan to expand Baby Steps Text to make keeping sentimental records as easy as sharing photos and videos of their child. The new version will allow parents to text memories via SMS or send photos and videos via MMS with the keyword “memory” to the system. Parents who have access to the Internet can view their records on a companion website alongside developmental reports for a complete picture of their child’s life.
7.2. General Implications for Text-Based mHealth
From our study, we found that extra keywords explained only on the instruction sheet (REVISIT, REPORT, HELP, etc.) were rarely used, except for one “HELP” and one “STOP” message. We had to ask participants to use the REPORT feature in the post-study survey to get their feedback on that feature. Therefore, we recommend to make all information inclusive in outgoing messages. Another suggestion is to allow adjustment of message frequency based on user preference if possible (immediately vs. periodically). System designers may also consider adding non-text features, such as web links or photos, to extend capabilities for users with Internet access to read detailed explanations that might help understanding or features that improve engagement (e.g., online photo albums). Future studies will explore these features and how technology interventions might “gracefully degrade” based on financial constraints. Next, shortening an already validated screening tool (e.g., ASQ) to 160 characters was necessary for the tool to function, but may have broken its validity. As using SMS (160-character limit) for health interventions becomes more common, we encourage creators of screening questionnaires to consider making questions fit within 140 characters or less to accommodate these constraints and to allow for additional characters for any instructions. Finally, communicating the results of a screener entirely via text message could create unnecessary anxiety [4] if done without including actionable information (such as calling a hotline) or without acknowledging the user’s feelings. Careful testing of wording in the results with potential users can help mitigate the potential for anxiety. Finally, our Wizard-of-Oz prototype was very useful in identifying problems with the messaging and syntax and required no technical skill to implement, as the only components were the scripts and the use of Google Voice. We suggest designers consider adopting this method for designing SMS-based applications.

7.3. Limitations & Future Work
Due to 160-character limit of SMS, we needed to shorten a validated screening tool and thus the validity of content needs to be reevaluated. We also did not follow up with the family health hotline to determine if any of the participants called to request access to services for early intervention, and only tested an English-language version of the tool. In the next iteration of Baby Steps Text, we plan to use a shorter validated screening tool that fits within character constraints (http://theswwy.org) and provide support in Spanish. Also, as this was an initial feasibility study, participants used Baby Steps for a brief period and we do not have sufficient data to tell how our system impacts parents in a long term (e.g. how it changes parental efficacy, their knowledge on child development, etc.) Therefore, we plan to conduct a longer study to answer questions about long-term engagement, identify ways to combine developmental tracking with sentimental record-keeping, and compare it to other approaches to technology-based developmental screening.

8. Conclusion
The primary objective of this study was to design and evaluate the feasibility of using SMS for childhood developmental screening. We presented our design process and implementation details for Baby Steps Text, an automated text messaging system to collect and screen child development using everyday technology available to resource-constrained populations. Our initial month-long deployment study shows that text messaging is a feasible tool for supporting parents in monitoring and screening their child. Continued use of our system beyond the study participation also indicates that the approach is promising in encouraging parent engagement.

References

2005
18. Jeong HY, Park SY, Zimmerman J. Opportunities to support parents in managing their children's health. CHI 2008 Extended Abstract (pp. 3225-3230). ACM.
20. Khokhar A. Short text messages (SMS) as a reminder system for making working women from Delhi Breast Aware.
23. Kientz JA. Understanding parent-pediatrician interactions for the design of health technologies. IHI 2010. ACM.
42. Yun TJ, Jeong HY, Hill TD, et al. Using SMS to provide continuous assessment and improve health outcomes for children with asthma. IHI 2012 (pp. 621-630). ACM.
Evaluating a Novel Summary Visualization for Clinical Trial Reports: A Usability Study

Maurine Tong, BS, William Hsu, PhD, Ricky K. Taira, PhD
University of California Los Angeles, Los Angeles, CA

Abstract
Contributions of clinical trials are captured in published reports that are unstructured and often require extensive manual review to gain a deeper understanding of the study itself. Our goal is to increase comprehension and decrease the time necessary to understand these reports through the use of visualization tools. In this paper, we specify and evaluate the visualization of a previously developed representation as well as gain insight from user input for further development. The usability experiment consisted of a two-arm study with users either having or not having access to the visualization. A user questionnaire was used to measure time spent and accuracy in comprehension; intuitiveness and reproducibility of the visualization; and preferences. We found that having the visualization required on average 28.1% less time (25.8 min vs. 35.8 min, p=0.01) while maintaining similar accuracy (73.7% vs. 67.0%). Users were then asked to create their own visualizations, with their visualizations averaging 86.1% similar to the gold standard. All participants either preferred the visualization over the status quo or preferred both equally. These results demonstrate that novel visualizations for trial reports could provide time savings and achieve similar accuracy as reviewing the paper itself. Understanding the strength and quality of clinical trials can be alleviated with a visualization that makes content explicit.

Introduction
Information within clinical trial reports can help inform clinical guidelines and support evidence-based medicine (EBM). Accurate interpretation of these results is important to characterize the quality and strength of a given clinical trial study. However, this information is difficult to assess, one reason is that its representation is free-text with numerical data scattered in tables, figures, and embedded in the text. A significant amount of effort is needed to identify and interpret information scattered throughout published reports, requiring the clinicians' and researchers' to organize this information mentally. This further complicates tasks such as assessing the strength of similar trials or comparing trials with conflicting results. The overarching goal of this work is to evaluate a visualization that links scattered data of a clinical trial together to better assess and understand its contributions.

The medical community is striving towards a structured representation to facilitate the deeper understanding of contributions from clinical trials. One attempt at standardizing the type of information within clinical trial reports is the Consolidated Standards of Reporting Trials (CONSORT), which comprises a checklist of essential items that should be included in reports of clinical trials and a diagram for documenting the flow of participants through a trial. CONSORT is a starting point to help investigators and others write or appraise trial reports, but its purpose is to promote good reporting, with no particular emphasis on appraising trial reports. Since CONSORT's founding, groups have worked on defining, structuring and standardizing information related to clinical trials. RCT Schema captures concepts related to a trial's design, basic intervention description, execution, administration, and results. These data structures create classes for specific types of information, with a standardized way to fill in classes and attributes. While structuring of data has been done mainly for the purpose of patient recruitment, few efforts have attempted to develop a representation for capturing the context of information for further interpretation. Clinical Research Eligibility Criteria Extraction and Representation (EliXR) addresses the disconnect in ambiguous eligibility criteria and clinical data results through developing a framework for eligibility criteria text. Our effort attempts to capture a wider spectrum of information and reported statistics necessary to perform quality assessment, relationships to piece information together, and a method for interactively presenting a published report to facilitate the user’s understanding of the study.

We have been developing a data model, which abstracts various types of information presented in clinical trial literature and connects these pieces together, and a visualization that allows target users (e.g., translational researchers, research students) to better interpret this information. We hypothesize that this representation and its visualization can facilitate understanding and analysis of clinical trial studies. In this paper, we specify a representation to classify information presented in clinical trial reports. We evaluate the visualization and report its performance, measured on a set of non-small cell lung cancer (NSCLC) clinical trial reports.
Methods

A usability study was performed to assess the benefit of using a standard visualization to understand the contents of a clinical trial report. In particular, we were interested in determining how well the visualization assists a researcher with understanding the published report's content, its ease of use, whether its presentation is intuitive to navigate and comprehend, and user preferences. Various aspects of the representation and visualization have been previously demonstrated\textsuperscript{10}, including methods for representing experimental flow\textsuperscript{11} and statistical analysis\textsuperscript{12}. In this paper, we briefly describe the representation model and the visualization. Results of the usability study were based on a comparison of interpreting information using the status quo published report versus the visualization.

The Representational Model

A representational model has been developed to store information presented in a published report. The first goal of this structured representation is to capture the essential elements related to recruitment, steps of the experiment, the data collection process, the data, the analyses, and the conclusions in a logical and consistent manner. The second goal is to provide context for observational data or reported statistics.

The model is functionally divided into purpose, methodology with raw data as a subgroup of methodology, statistical methods, and interpretations. These sections parallel the structure of a clinical trial report. Under the purpose section, the model captures the overall goal of the clinical trial with a free-text statement. The methodology section of the model contains a process model of nodes with a list of relevant variables, attributes, and values. Table 1 shows typical examples from a NSCLC clinical trial report\textsuperscript{13}.

Table 1. Example excerpts from clinical trial report\textsuperscript{13}.

<table>
<thead>
<tr>
<th>Excerpts from report</th>
<th>Node</th>
<th>Variables within Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Patients received up to six cycles of carboplatin/paclitaxel. Paclitaxel (200 mg/m2) was administered over 3 hours every 3 weeks. Carboplatin dosing was based on the Calvert formula with a target area under the curve of 6 mg/mL x min and glomerular filtration rate (GFR) estimated for males as GFR = (140-age) x weight/72 x (serum creatinine).”</td>
<td>Treatment</td>
<td>• drug name(s)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• dosage(s)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• frequency</td>
</tr>
<tr>
<td>“Eleven patients discontinued treatment as a result of a nonfatal AE. Discontinuations occurred as a result of: hemorrhagic event (three patients) in the low-dose bevacizumab arm; a hemorrhagic event (one patient); Aspergillus lung abscess (one patient); aspiration pneumonia (one patient); thrombotic stroke (one patient); vertebral fracture (one patient); and peripheral europathy (paclitaxel-related; one patient) in the high-dose arm. In two cases, bevacizumab was discontinued following initiation of anticoagulant therapy. Bevacizumab was withheld from one patient with subclavian vein thrombosis.”</td>
<td>Discontinued Treatment</td>
<td>• Hemorrhagic event</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Aspergillus lung abscess</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• aspiration pneumonia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• thrombotic stroke</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• vertebral fracture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• peripheral neuropathy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• following initiation of anticoagulant therapy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• subclavian vein thrombosis</td>
</tr>
<tr>
<td>“The overall response rate showed a trend toward improved response for patients receiving bevacizumab, with the highest response noted in the high-dose group (31.5%) and the lowest in the control group (18.8%; Table 2)....Median TTP was longer in the high-dose bevacizumab arm compared with the control arm (7.4 v 4.2 months; ( P = .023 ); Fig 1)....Survival for the high-dose bevacizumab arm was modestly longer than the control arm (17.7 v 14.9 months; ( P = .63 ); Fig 2).”</td>
<td>Survival</td>
<td>• median survival (month)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• median survival (range)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% PFS vs. time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% survival vs. time</td>
</tr>
</tbody>
</table>

The process model serves three purposes: (1) it documents the recruitment process using a flow chart to represent eligibility criteria, (2) demonstrates randomization by branching of nodes from one to many nodes, with each node representing a subpopulation, and (3) captures the overall study flow using nodes to represent specific types of experimental procedures. Typical types of nodes include general population, population sampling pool, decision boxes, recruitment criteria, control and intervention population, methods, and observation points. Each node is associated with a list of relevant variables, attributes, and values within a spreadsheet structure corresponding to the raw data stated in the published report. Together, the process model and linked spreadsheet allow fragmented knowledge pre-
sented in free-text, figures, and tables, to be placed in the context of the entire clinical trial "experiment." The statistical methods section of the model collects information from the text and displays it in a consistent manner with references back to the nodes of the process model and data elements from the spreadsheet. Each statistical method includes inputs from the spreadsheet, outputs as a measure of statistical significance, and a statement of significance. The interpretations section of the model captures the contributions of the clinical trial as stated by the published report.

The Visualization

The visualization assists with viewing and interacting with the contents of the representation from each clinical trial report. The visualization is divided into four panels following the structure of the model: (1) purpose of the trial, (2) process model and data grid, (3) list of statistical methods with variables compared and results, and (4) interpretations. The purpose of the trial and the endpoints are listed at the top of the visualization in free-text. The recruitment process and methodology are below. At the bottom are the interpretations summarizing the contributions of the published report listed in free-text (Figure 1).

The process of a clinical trial study is displayed as a timeline of events performed over the course of the entire study. In this example, the study population is recruited using three separate types of criteria: tumor status, presence of prior chemotherapy, and other clinical criteria. The study population is then randomized into three groups, control, low dose, and high dose; and each group undergoes a set of protocols. Note that the control receives no drug intervention as shown by a lack of yellow boxes in the control row (Figure 2).

Figure 2. Flow chart of events.

Directly below the diagramming interface, the methodology, raw data, and results can be viewed as a data grid. The variables appear on the left side, with one variable per row. Thus, the rows give an inventory of variables that are mentioned in the clinical trial study. The columns of the data grid correspond to different nodes in the process model. Any numerical data generated is placed along the same row as the variable and is beneath its corresponding event. Thus, each cell in the data grid is associated with an event node from the flow diagram and variable and the cell itself corresponds to the specifications or characterization of a variable for an experimental procedure of a group of patients for which the node refers to in the flow chart. A cell can show (1) summary statistics, and (2) all the patient values for a given variable, if available (Figure 3). We are working towards the customization of visualizing the data from individual cells. We are also developing methods to derive useful information from the data within related cells (e.g., points in Kaplan-Meier diagram where curves cross).

Figure 3. (Left) Example of data stored within a cell. Data is sorted by population and organized as individual data or population statistics. (Right) Graph generated by data. In this example, a Kaplan-Meier graph is drawn.

The panel for statistical methods provides a visual inventory of all the tests performed. Each test is listed with its corresponding inputs from the data grid, the test statistic, output statistics such as a p-value, and a statement of significance.
significance (Figure 4). For the purpose of the usability study, participants were presented with a mockup of the visualization in Microsoft Visio for each published report created; the final version is being implemented as an interactive application using Java.

Figure 4. Statistical method visualization. Inputs include experimental arm name, sample size, and variable. This is followed the statistical test within the box. Outputs include a quantitative measure of statistical significance, and a statistical statement.

Overview of Usability Study

The usability study had three objectives: (1) test the ease of interpretation and time saved against the status quo (i.e., the published report), (2) test the intuitiveness of the visualization through reproducibility of the visualization, and (3) elicit preferences from a targeted user group.

Paper cohort: Clinical trials were randomly chosen based on a PubMed search using the keywords "EGFR", "lung cancer", "non-small cell lung cancer", and "clinical trial" (Table 2). The search yielded 38 published reports which were different from the set of reports used to develop the original visualization. For the initial scope of this study, we randomly selected three papers that met the criteria of being a clinical trial about NSCLC involving EGFR mutations to assess time spent and accuracy. The fourth report was used for the study on reproducibility.

Table 2. Randomly selected clinical trial reports on non-small cell lung cancer.

<table>
<thead>
<tr>
<th>Trial number</th>
<th>Title of Report</th>
<th>Outcome variable</th>
<th>Sample size</th>
<th>Total Events</th>
<th>Date Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Randomized Phase II Trial Comparing Bevacizumab Plus Carboplatin and Paclitaxel With Carboplatin and Paclitaxel Alone in Previously Untreated Locally Advanced or Metastatic Non-Small-Cell Lung Cancer13</td>
<td>Response Rate (RR)</td>
<td>99</td>
<td>33</td>
<td>2005</td>
</tr>
<tr>
<td>2</td>
<td>First-Line Gefitinib in Patients with Advanced Non-Small-Cell Lung Cancer Harboring Somatic EGFR Mutations14</td>
<td>Objective Response Rate (ORR) = sum of patients with confirmed complete and partial responses / number of patients treated</td>
<td>98</td>
<td>15</td>
<td>2008</td>
</tr>
<tr>
<td>3</td>
<td>EGFR expression as a predictor of survival for first-line chemotherapy plus cetuximab in patients with advanced non-small-cell lung cancer: analysis of data from the phase 3 FLEX study15</td>
<td>Overall Survival (OS)</td>
<td>1125</td>
<td>24</td>
<td>2011</td>
</tr>
<tr>
<td>4</td>
<td>First-Line Gefitinib in Patients with Advanced Non-Small-Cell Lung Cancer Harboring Somatic EGFR Mutations16</td>
<td>Objective Response Rate (ORR) = sum of patients with confirmed complete and partial responses / number of patients treated</td>
<td>889</td>
<td>17</td>
<td>2008</td>
</tr>
</tbody>
</table>

Participant recruitment: The participants were students with familiarity in biostatistics and informatics. In this usability study, students served as a proxy for potential users (i.e., biostatisticians and clinician researchers). The study for time spent and accuracy, and for eliciting preferences had 11 participants, and the study for reproducibility had 6 participants. All participants were asked in a pre-questionnaire about the length and fluency of their experience with clinical trial reports, non-small cell lung cancer, and statistics.

Task definition to assess time spent and accuracy: For the first objective, the participants used the visualization to answer questions demonstrating their comprehension of the clinical trial and recorded the time required to answer each question. Questions were divided into two types: comprehension to assess whether the individual is able to synthesize evidence from the published report, and information retrieval (IR) to focus on locating specific pieces of evidence in the report. Comprehension questions were developed based on reporting guideline requirements5, and
asked readers to interpret the objective and claims made in the published report. For example, one comprehension question asked: "The trial states 'This large prospective biomarker study found that patients with activating EGFR mutations derive the greatest PFS benefit from erlotinib maintenance therapy.' Describe the method, numerical data, and analyses for this statement." IR questions focused on locating key information as adapted from applicable CONSORT requirements. IR questions include reporting the eligibility criteria, locating the experimental arms, summarizing the methodology, and identifying the results of statistical tests. Questions of both types were presented using multiple choice, fill-in-the-blank, and short answer. All questions were reviewed by a biostatistician who was not involved in the development of the system to reduce bias in word-choice and to ensure conformance to standard guidelines and terminology. The gold standard was created by a domain expert who was given an open amount of time. Tasks were timed and graded for accuracy by determining the percentage of questions answered correctly.

**Study design to assess time spent and accuracy:** A two-arm randomized trial design was used with 11 participants and 3 clinical trial studies. The presentation of the trial was randomized for each participant and each participant reviewed clinical trial 1, 2 and 3 (Table 2). For each clinical trial, participants were randomized into the visualization study arm or the status quo study arm. In either study arm: (1) participants filled out paperwork (consent form, pre-test questionnaire) and received a tutorial on how to interpret the visualization based on two example questions; then (2) participants completed the usability sessions either with the status quo or visualization; finally (3) participants answered a post-questionnaire about the visualization (Figure 5).

**Task definition to assess reproducibility:** For the second objective, the participants were asked to generate a section of the visualization (i.e., the experimental flow) using a provided set of guidelines and a tutorial. Reproducibility was assessed through percent match of elements, nodes and variables. Elements are annotations given to a node, and can include sample size and time points. Nodes are used to label processes for recruiting a sample population and intervention processes. Variables describe and characterize a node. Example variables include randomization methods, baseline experimental procedures, and survival assessment methods. Tasks were compared with a gold standard created by a domain expert who was given an open amount of time.

**Task definition to assess user preferences:** A post-questionnaire was given to assess the affinity and usefulness of the visualization to gather impressions on the adequacy of its contents and to provide feedback on design, interface, and suggestions for additional functionalities. Participants answered questions using a Likert scale from one to ten, ten being very satisfied and very usable.

**Statistical Methods**

The participants used the status quo or the visualization to answer questions demonstrating their comprehension of the clinical trial and recorded the time required to answer the questions. The dependent measures of this usability study included time spent, measured in minutes; and accuracy, measured as the percentage of questions answered correctly. A weighted accuracy, combining time and accuracy, was calculated to place more emphasis on high accuracy scores that were obtained in a shorter amount of time. Descriptive statistics were computed for all measures.

\[
Weighted\ Accuracy = \frac{1}{Time} \cdot Accuracy
\]

Overall time spent and accuracy was determined by averaging over all values in each condition. Groups were conditioned on having either the status quo paper representation or the visualization. A pilot study was used to estimate the amount of time and accuracy for each task that was considered reasonable. A power calculation was performed to determine the appropriate sample size for the combination of participants and clinical trials needed. With an estimated time difference of 10 minutes (30 minutes vs. 40 minutes) and standard deviation of 8 minutes, a sample size of 12...
per group would yield an 83% power with 5% significance level. With an estimated accuracy difference of 15% (70% vs. 85%) and standard deviation 17%, a sample size of 12 per group would yield an 80% power with 5% significance level. Hence, a sample size of at least 24 is needed, meaning at least 8 participants each reading 3 clinical trial reports. This is satisfied by the number of participants enrolled. A 2-sided student’s t-test was used to compare (1) accuracy, (2) time spent, and (3) weighted accuracy using the visualization vs. using the status quo method.

Results

Participant Characteristics: Eleven student participants were involved in the study for time spent and accuracy, and for eliciting preferences; and six additional student participants were involved in the study for reproducibility. Participants ranged in experience from one to five years participated. All participants have read a clinical trial report before and took on average 80 minutes to read it completely. General participant characteristics are presented in Table 3.

Table 3. Characteristics of Participants. For confidence, scale values are 1 = not confident and 10 = very confident. For courses, values indicate number of college/graduate level courses.

<table>
<thead>
<tr>
<th></th>
<th>Time &amp; Accuracy</th>
<th>Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=11)</td>
<td>(n=6)</td>
</tr>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Confidence with understanding of cancer mechanisms</td>
<td>5.09 ± 1.868</td>
<td>4.67 ± 3.055</td>
</tr>
<tr>
<td>Confidence with knowledge on NSCLC</td>
<td>4.36 ± 1.963</td>
<td>4.33 ± 3.055</td>
</tr>
<tr>
<td>Confidence in understanding knowledge within reports</td>
<td>6.09 ± 1.921</td>
<td>6.33 ± 1.528</td>
</tr>
<tr>
<td>Confidence in understanding statistical methods</td>
<td>6.36 ± 1.502</td>
<td>2.00 ± 1.000</td>
</tr>
<tr>
<td>Confidence in assessing the quality of statistical tests</td>
<td>4.45 ± 1.916</td>
<td>5.33 ± 3.215</td>
</tr>
<tr>
<td>Courses on biology-undergraduate</td>
<td>3.18 ± 2.676</td>
<td>6.67 ± 2.887</td>
</tr>
<tr>
<td>Courses on biology-graduate</td>
<td>3.09 ± 5.665</td>
<td>2.33 ± 1.528</td>
</tr>
<tr>
<td>Courses on statistics-undergraduate</td>
<td>1.27 ± 0.786</td>
<td>7.67 ± 2.082</td>
</tr>
<tr>
<td>Courses on statistics-graduate</td>
<td>2.27 ± 1.191</td>
<td>3.00 ± 2.000</td>
</tr>
</tbody>
</table>

Time spent and accuracy: In this usability study, time spent and accuracy was measured using self-reported completion time, and percent questions answered correctly. Non-significant differences were found in both time and accuracy between the visualization condition and the status quo condition for each clinical trial study (Table 4). The point estimate of report #2 was shown to have decreased accuracy as compared with report #1 and #2. The accuracy can be affected due to an increase in complexity of the study design and greater amount of content for both the visualization and status quo method. The accuracy for comprehension question and for IR questions were separated for exploratory analyses. The mean accuracy for comprehension questions within one reports suggests a difference between the visualization condition and the status quo condition, favoring the visualization (data not shown). This suggests that using the visualization can increase comprehension. This trend within reports are currently being studied in an attempt to significantly increase accuracy in the visualization.

Overall accuracy was similar between the visualization and status quo, however, participants with the visualization had on average a quicker overall time than participants with the status quo (visualization 25.8 ± 10.10 minutes vs. status quo 35.8 ± 9.94 minutes; p<0.01). This suggests that information is easier to locate in a visualization than in the status quo. While the visualization provided similar accuracy, the tradeoff is a significant times savings when compared to the status quo alone. The weighted accuracy results further demonstrated that participants with the visualization had a combined quicker time and increased accuracy than participants with the status quo (visualization 3.39 ± 1.66 vs. status quo 2.10 ± 0.87, range=0.71 to 7.00, p<0.008).

Table 4. Measures of performance as a function of errors and time.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Visualization</th>
<th>Status Quo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy (%)</td>
<td>SD</td>
</tr>
<tr>
<td>1</td>
<td>77.3%</td>
<td>5.34%</td>
</tr>
<tr>
<td>2</td>
<td>58.5%</td>
<td>11.67%</td>
</tr>
<tr>
<td>3</td>
<td>84.5%</td>
<td>6.10%</td>
</tr>
<tr>
<td></td>
<td>68.9%</td>
<td>10.01%</td>
</tr>
<tr>
<td></td>
<td>53.8%</td>
<td>14.80%</td>
</tr>
<tr>
<td></td>
<td>78.8%</td>
<td>10.23%</td>
</tr>
</tbody>
</table>

Reproducibility: Compared with the gold standard, the visualizations created by the six participants with limited training were 86.1% similar to the gold standard with a standard error of 6.45%. Dissimilarity was due to incomplete
visualizations as opposed to visualizations that differ fundamentally in structure. Out of all errors, 32% of the dissimilarity was due to neglecting to fill in available time points; 24% was neglecting to put in baseline experimental procedures; 12% was neglecting to fill in survival assessments time points; 12% was neglecting to put in events that can change the sample size, such as drop outs; and 20% was due to neglecting to fill in sample sizes for populations. The reasons for error and the high similarity score shows that our representation is logical and easy to understand for the participants in our usability study.

Preferences: All eleven participants preferred the visualization to the status quo, or preferred both the report and the visualization. No one preferred the status quo to the visualization. Participants rated the usefulness on average as 8.0 (7.1-8.9, where 10 is completely essential), and the satisfaction of the visualization at the current state as 7.8 (6.7-8.9, where 10 is completely satisfied). The likelihood of participants using the visualization again is reported on average as 87.2% with a standard error of 10.7%. Additional comments mentioned participants appreciated the clean and well-formatted visualization, but wanted more context, such as information covered in the introduction and discussion. Participants suggested additional development and/or functionality, such as having a summary box to better navigate a visualization that contains too many details and including the ability to distinguish significant findings visually. In general, participants found the visualization to be a good overview that assisted in understanding the clinical trial study.

Discussion

Understanding the strength and quality of clinical trials is a critical step in providing better healthcare to medical practice. Our previously developed visualization places essential information in context with a process model. A user can navigate through nodes to identify the population involved, its sample size, the procedures performed, and data that result. Because the visualization can help summarize essential elements and connect relevant elements together, it can be valuable for biostatisticians and research clinicians, who routinely access information from within clinical trial reports, and are considered our potential users. With any system, it is critical to examine usability issues pertaining to people of varying backgrounds in academia successfully accessing information needed within clinical trial reports. The goal of this study was to determine whether our previously designed visualization could be used to increase time saved and accuracy as compare to the current method, while maintaining an easy to understand format.

The results of the usability study were consistent with our intuition. We found that having the visualization required on average 28.1% less time (25.8 min vs. 35.8 min; p=0.01) while maintaining similar accuracy. These findings did not appear to be affected by participants’ varying levels of familiarity with the statistics, clinical domain (i.e., non-small cell lung cancer) and clinical trial procedures. This suggests that having essential information placed in context of the entire experiment helps users cognitively critique and apply contributions of clinical trials on a deeper level in a more timely fashion. This enables informatics tools to query information to be used for meta-analysis and probabilistic disease modeling and assist with the difficult task of assessing the quality and usefulness of each trial. The results of the reproducibility study showed that visualizations were on average 86.1% similar to the gold standard when produced by participants. This suggests that the proposed visualization is easy to understand and apply, and logically represents essential elements in a clinical trial study. This paper presents an attempt at identifying a standardized view that follows from intuition given that users have formal training in informatics.

While all participants favored the visualization over the current method, questionnaires revealed that much work is needed to improve the satisfaction and usability of the visualization. One solution to avoid bias of a less completely documented clinical trial study is to use the visualization to supplement an individual’s understanding gained from reading the status quo published report. While the study design proposed in this paper assigns participants to either the status quo or the visualization condition, in actuality, the two conditions are not mutually exclusive. This suggests that the combination of having a visualization to reference while reading the status quo published report can further help to save time and increase accuracy. One of the next steps include extending the evaluation to potential users. While the results from this study were gathered on students serving as surrogates, preliminary interviews with potential users show promise. In an unstructured interview, one biostatistics professor anecdotally noted that she liked the hybrid process model-spreadsheet for contextualizing observations and statistics.

We recognize one limitation is the current reporting of clinical trial studies, which may contain missing information. Because the visualization is designed to present the same information as the status quo, missing information can negatively affect the user satisfaction rating. Another limitation is related to the design of the task questionnaires. Because no standard list of questions exist to test comprehension of clinical trials, we tailored questions from standard reporting guidelines to determine the types of information necessary for comprehension. The final list of questions was confirmed by domain experts to determine if answering questions display understanding. We also need to address
the limitation of self-reporting (e.g., timing), which may be inaccurate. Another limitation includes the recruitment of participants and the selection of baseline assessment measures in the pre-questionnaire analysis. Participants were students that served as proxy for potential users. Classes provided a quantitative way to rank students, however, the assessment may not be accurate for non-traditional students who gained experiences outside of classes or have a time-lapse of many years between graduate and undergraduate courses. One more limitation relates to the issue with questionnaires based on the Likert scale. While Likert scales can pinpoint problem areas, they are unable to give information on the nature of the identified problem.

Conclusion

The current research addresses the usefulness of the visualization and tests its efficacy in understanding clinical trials in a timely manner. Our results suggest that the application can decrease time without sacrificing accuracy, the visualization is reproducible across multiple users, and the system is generally accepted in a targeted user group. With the abundance of clinical trial research, further work is necessary to translate the published report into a concise and informative visualization containing the same information but with more functionality. The logical structure of the results of clinical trial reports will allow computers to better assess the quality and strength of trials, extract important statistics, and assist in comparison of trials.

Acknowledgements

This work is supported by the NIH grant number LM011333. We thank James Sayre, PhD., and Grace Hyun Kim, PhD., for their helpful discussions and constructive criticisms. We also thank our domain experts, our pilot study participants, and all our usability study participants.

References

10. Tong M, Hsu W and Taira RK. A representation for standardizing numerical data from clinical trial reports. 2012 RSNA Scientific Assembly and Annual Meeting Bioinformatics Exhibit. LL-INE1228-TUA


Nurse Informaticians Report Low Satisfaction and Multi-level Concerns with Electronic Health Records: Results from an International Survey

Maxim Topaz PhD, MA, RN¹, Charlene Ronquillo, MSN, RN², Laura-Maria Peltonen, MNSc, RN³, Lisiane Pruinelli, MSN, RN⁴, Raymond Francis Sarmiento, RN⁵, Martha K. Badger, MSN, RN-BC, CPHIMS⁶, Samira Ali, MSN, RN⁷, Adrienne Lewis, MSc, MsN(c), RN⁸, Mattias Georgsson, MSc, RN⁹, Eunjoo Jeon, RN¹⁰, Jude L. Tayaben, MAN, RN¹¹, Chi-Hsiang Kuo, RN¹², Tasneem Islam, RN¹³, Janine Sommer, RN¹⁴, Hyunggu Jung, RN¹⁵, Gabrielle Jacklin Eler, RN¹⁶, Dari Alhuwail, RN¹⁷, Ying-Li Lee, MSN, RN¹⁸

¹Harvard Medical School & Brigham and Women's Hospital, Boston, USA; ²University of British Columbia, Vancouver, Canada & University of the West of England; ³Nursing Science, University of Turku and Turku University Hospital, Turku, Finland; ⁴University of Minnesota, School of Nursing, MN, USA; ⁵National Institute for Occupational Safety and Health, U.S. Centers for Disease Control and Prevention, USA; ⁶University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA; ⁷Carlow University, Pittsburgh, USA; ⁸Independent Researcher, Canada; ⁹Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden; ¹⁰College of Nursing, Seoul National University, Seoul, Republic of Korea; ¹¹College of Nursing, Benguet State University, La Trinidad, Benguet, Philippines; ¹²Tzu Chi University of Science and Technology, Taiwan; ¹³Deakin University, Victoria, Australia; ¹⁴Hospital Italiano, Buenos Aires, Argentina; ¹⁵Biomedical and Health Informatics, University of Washington, Seattle, USA; ¹⁶Instituto Federal do Paraná, Londrina, Brazil; ¹⁷Department of Information Systems, College of Engineering & Information Technology, University of Maryland, Baltimore County, USA; ¹⁸Institute of Biomedical Informatics, National Yang Ming University, Taipei, Taiwan

Abstract
This study presents a qualitative content analysis of nurses’ satisfaction and issues with current electronic health record (EHR) systems, as reflected in one of the largest international surveys of nursing informatics. Study participants from 45 countries (n=469) ranked their satisfaction with the current state of nursing functionality in EHRs as relatively low. Two-thirds of the participants (n=283) provided disconcerting comments when explaining their low satisfaction rankings. More than one half of the comments identified issues at the system level (e.g., poor system usability; non-integrated systems and poor interoperability; lack of standards; and limited functionality/missing components), followed by user–task issues (e.g., failure of systems to meet nursing clinical needs; non nursing-specific systems) and environment issues (e.g., low prevalence of EHRs; lack of user training). The study results call for the attention of international stakeholders (educators, managers, policy makers) to improve the current issues with EHRs from a nursing perspective.

Introduction
The benefits of converting from paper-based systems to electronic health records (EHRs) are being advocated for in various nations around the globe, leading for a push for healthcare professionals, including nurses, to adopt EHRs. In 2012, the EHR adoption rates were relatively high and those numbers grew in the last few years, for example: Sweden and Germany adoption rate was over 80%, the United States (U.S.) 69%, France 67%, Canada 56%, and Switzerland 41%¹.

An increased awareness of the importance of usability and other system issues have accompanied the increased adoption of EHRs by healthcare providers². A systematic review of empirical studies of EHRs identified the potential benefits of this technology in supporting patient care and clinical documentation³. However, reaping these benefits require addressing challenges related to implementation, adoption, and satisfaction of EHRs. The careful consideration of sociotechnical contexts and links between the clinicians, patients, and technology should be taken into account during the EHR development and implementation processes⁴. Barriers to EHR adoption included cumbersome system functionalities, lack of interoperability, and hardware issues in a study examining the use of EHR in community settings in the U.S.⁴. This study highlighted the impact of usability on nurses’ workflow, satisfaction, efficiency, and adoption of the EHR, as well as the importance of considering the interaction between system functionality, usability, and clinician workflow. The far reaching
negative effects of poor system usability are exemplified by the call for solutions to poorly designed EHRs by the American Medical Association in 2014. This resulted from physicians’ frustration and dissatisfaction with EHR usability and increasingly negative perceptions of EHRs since 2010. Nurses have expressed similar dissatisfaction with EHRs related to the system’s poor fit with clinical workflow, disruptions to productivity, and negative impacts on nurse-patient communication. In one survey of 13,650 nurses in the U.S. in 2014, 92% expressed dissatisfaction with EHRs, 85% noted that they struggled with continually flawed systems, and 67% reported using workarounds to avoid the unresolved flaws of EHR systems. Beyond causing frustration and being time-consuming for clinicians, poor EHR usability may also result in adverse events, medical errors, and other unintended negative consequences. Clearly, the usability of EHRs has important implications for patient safety and quality of care.

While the impact of EHR implementation on nursing is increasingly recognised, the complexities of how the technology is adopted in clinical nursing practice remains poorly understood. Nevertheless, nurses are still called upon to be key drivers towards the move from paper-based to electronic systems. Furthermore, nurses’ participation in decision-making, development, and evaluation of EHR development and implementation is being increasingly emphasized. Although nurses represent approximately one-third of hospital employees and nursing is one of the largest EHR user groups, nurses’ perceptions of EHRs are rarely surveyed and remain largely unknown. For example, only two out of 346 identified usability studies of healthcare information technology conducted between 2003 to 2009 examined systems use by nurses. A specific examination of usability issues related to nurses’ use of EHRs is warranted as the usability issues faced by nurses may differ from the issues faced by other healthcare professions.

The International Medical Informatics Association–Nursing Informatics Special Interest Group (IMIA-NISIG) Students Working Group members aimed to fill this gap in health information technology research by conducting an international survey of the state of nursing informatics. Our survey solicited responses from participants from 45 countries on various health informatics topics, one of which was the respondents’ perceptions of the usability of EHR systems. In this paper, we present the results of two survey questions related to respondents’ level of satisfaction with, and comments relating to, the current state of EHRs used by nurses.

Methods

Survey creation and distribution

This study had a cross-sectional survey design with online data collection. The questionnaire was developed based on current nursing informatics literature that explored current and future trends in nursing informatics. The questionnaire was iteratively developed, revised, and edited by the members of the IMIA-NI students working group. The group also shared the questionnaire with several international nursing informatics experts within the IMIA-NISIG leadership. Experts were defined as individuals with multiple publications that examined general informatics trends. Based on several rounds of expert recommendations, comments, and feedback, the group revised the questionnaire until a final version was developed. The online survey version used Google forms and was pilot-tested to assure its adequate functionality before the international distribution.

The study received a supportive ethical statement (Institutional Review Board exempt approval) from the University of Turku (Finland), where it was coordinated. The questionnaire was translated from English into six languages (Arabic, Korean, Portuguese, Spanish, Mandarin and Swedish). Each translation was conducted by a native speaking nurse with a background in informatics. The translation was validated by at least two other native speaking nursing informatics professionals and revised until the final version for distribution was generated. Responses to the open ended questions collected with the different translated versions of the survey were translated back to English for analysis. The translations were conducted by the nurses who translated the original survey questions and each translated response was validated for accuracy by one or two additional native language speakers fluent in English. Data were collected in August - October 2015. The following inclusion criteria were outlined in the survey invitation: any nurse (or other allied health professional) with experience in nursing informatics either in clinical practice or academia was eligible to participate. These groups were targeted in an effort to obtain a comprehensive overview of current trends in academia, as well as to explore issues with EHRs identified by clinicians. We used snowball sampling strategy to reach as many international respondents as possible. The IMIA-NISIG student working group’s members were invited to collaborate on the study and distribute a cover letter and the links to the survey to their networks. Eighteen students from fourteen countries actively participated in distributing the survey through their professional networks. These included global and local health informatics associations (e.g., IMIA, AMIA, etc.), clinical settings (hospitals and outpatient settings), and academic institutions.

The survey consisted of 24 questions with both structured and open-ended response options. Eight of the questions were focused on demographics (i.e., professional background; highest degree received; clinical or academic position; years of
informatics experience; and country and city) and eighteen questions pertained to the current or future state of nursing informatics.

To date, results of two survey questions related to current and future trends in nursing informatics have been presented in two published papers. This paper focuses on the two questions that related to respondents’ satisfaction with the current state of EHRs used by nurses. The two questions were: 1) Are you satisfied with the current state of nursing computerized documentation (in electronic health records) in your country/hospital?; and 2) If you are not satisfied with the current state of nursing computerized documentation (in electronic health records) in your country/hospital, please provide a few reasons. Responses to question one were collected using sliding Likert scale ranging from 1 to 10, where 1 indicated the lowest level of satisfaction and 10 the highest level. Responses to question two were open ended without a text length limitation.

**Conceptual framework:** A Stratified View of Health Information Technology usability evaluation (SV-HIT) was used as a conceptual framework to guide the qualitative analysis in this study. SV-HIT was created as a result of a review of 319 articles that evaluated health information technology used by clinicians. The model has 3 levels, as illustrated in Figure 1.

**Level 1- User-Task Issues:** Targets system specifications to understand any issues with user–task interactions during the system development. This level focuses on several key questions, including “How can an information technology system be used to support the needs/tasks?” and “What are the user needs?”; **Level 2- System Issues:** Targets task performance to assess systems and human–computer interaction, user–task–system or system–task interactions. The key questions are: “Does the information system work for the task?” and “What is the quality, speed, accuracy, and completeness of user -system interaction performance?”; **Level 3- Environment Issues:** Incorporates environmental factors to identify work processes and system impact in a real clinical setting. Key questions include: “What are the organizational, user and other factors affecting system use?” and “How does the system impact healthcare?”

**Quantitative analysis:** First, the general sample demographic and the professional characteristics were summarized. Then, respondents’ characteristics were compared between those who provided comments on issues with EHR versus those who did not, using standard statistical tests (Chi-square/exact Fisher test or t-test).

**Qualitative analysis:** Open-ended questions were independently analysed by two authors experienced with qualitative data analysis (CR and MT). Thematic analysis, a qualitative descriptive approach for identifying, analysing and reporting themes within data, was used to analyse the responses. First, each response was examined independently by the two authors and one or more themes for each response were suggested. Data were collated in an Excel spreadsheet. For example, the response, “The current EHR systems are cumbersome, not intuitive, and require too much clicking,” was classified as related to “Poor usability”. After an initial categorization of all the responses, the authors discussed the themes that emerged and consolidated them into eight major themes and an additional Other category. Each author then went back to the original responses and revised the themes for each of the responses. The themes for each response were then merged and the two authors achieved consensus on the themes for each response. Finally, each theme was mapped to one of the 3 SV-HIT levels (task- user/ system/ or environmental issues). For example, Poor usability theme was mapped to the System issues level whereas Lack of support at the hospital or policy levels theme was mapped to Environmental issues level. The results were shared with two additional authors for validation. The summary statistics and description of the major themes that emerged are presented in this paper. As illustrated in Figure 1, the three different levels have some degree of overlap. This was evident in our work where some themes that emerged were inter-related. For example, the Limited system functionality theme (System issues level) was related to the Systems failing to meet nursing clinical needs theme (User-task level), and vice versa.

![Figure 1: A Stratified View of Health Information Technology usability evaluation (SV-HIT), adapted from](image-url)
Results

A total of 469 respondents (the sample for this paper) answered the question regarding their satisfaction with EHRs used by nurses. Table 1 presents the sample characteristics. Most of the participants were nurses (89%). The rest were other healthcare professionals, including health informaticians, pharmacists, biomedical engineers, etc. Most of the participants had at least a Master’s degree and more than five years of experience in informatics. Slightly more than one-third (36.2%) of the participants had formal training in informatics. When the question was relevant, most of the participants indicated that they occupied a middle management position or higher and/or held an academic rank of teacher or professor.

Table 1: Survey participant characteristics

<table>
<thead>
<tr>
<th>Category</th>
<th>Total study population (n=469)</th>
<th>Respondents who commented on issues with EHR (n=283)</th>
<th>Respondents who did not comment on issues with EHR (n=186)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Profession</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nurse</td>
<td>420 (89.6%)</td>
<td>255(90.2%)</td>
<td>165(88.7%)</td>
</tr>
<tr>
<td>Other</td>
<td>44(9.4%)</td>
<td>27(9.5%)</td>
<td>17(9.1%)</td>
</tr>
<tr>
<td>Education**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelors</td>
<td>136(29%)</td>
<td>67(23.5%)</td>
<td>69(37.1%)</td>
</tr>
<tr>
<td>Masters</td>
<td>201(43.3%)</td>
<td>117(41.8%)</td>
<td>84(45.2%)</td>
</tr>
<tr>
<td>PhD</td>
<td>111(23.5%)</td>
<td>88(30.9%)</td>
<td>23(12.4%)</td>
</tr>
<tr>
<td>Other</td>
<td>20(4.3%)</td>
<td>11(3.9%)</td>
<td>9(5.4%)</td>
</tr>
<tr>
<td>Years of nursing informatics experience**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 years</td>
<td>178(38%)</td>
<td>94(33%)</td>
<td>84(45.2%)</td>
</tr>
<tr>
<td>6-10 years</td>
<td>100(21.3%)</td>
<td>57(20%)</td>
<td>43(23.1%)</td>
</tr>
<tr>
<td>11-15 years</td>
<td>69(14.7%)</td>
<td>45(15.8%)</td>
<td>24(12.9%)</td>
</tr>
<tr>
<td>16-20 years</td>
<td>54(11.5%)</td>
<td>40(14%)</td>
<td>14(7.5%)</td>
</tr>
<tr>
<td>21-45 years</td>
<td>51(10.9%)</td>
<td>40(14.7%)</td>
<td>11(5.9%)</td>
</tr>
<tr>
<td>Formal training in informatics*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>170(36.2%)</td>
<td>105(37.2%)</td>
<td>65(34.9%)</td>
</tr>
<tr>
<td>No</td>
<td>267(56.9%)</td>
<td>152(53.7%)</td>
<td>115(61.8%)</td>
</tr>
<tr>
<td>Other</td>
<td>28(6%)</td>
<td>23(8.1%)</td>
<td>5(2.7%)</td>
</tr>
<tr>
<td>Clinical Position (if relevant)**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff</td>
<td>128(27.3%)</td>
<td>62(21.8%)</td>
<td>66(35.5%)</td>
</tr>
<tr>
<td>Middle management</td>
<td>133(28.4%)</td>
<td>77(27%)</td>
<td>56(30.1%)</td>
</tr>
<tr>
<td>Upper management</td>
<td>55(11.7%)</td>
<td>42 (15.1%)</td>
<td>13(7%)</td>
</tr>
<tr>
<td>Other</td>
<td>76(16.2%)</td>
<td>56(20%)</td>
<td>20(10.8%)</td>
</tr>
<tr>
<td>Academic Position (if relevant)**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>75(16%)</td>
<td>36(12.6%)</td>
<td>39(21%)</td>
</tr>
<tr>
<td>Teacher</td>
<td>73(15.6%)</td>
<td>34(11.9%)</td>
<td>39(21%)</td>
</tr>
<tr>
<td>Professor</td>
<td>91(19.4%)</td>
<td>73(26%)</td>
<td>18(9.7%)</td>
</tr>
<tr>
<td>Other</td>
<td>58(12.3%)</td>
<td>44(15.8%)</td>
<td>14(7.5%)</td>
</tr>
<tr>
<td>World Health Organization (WHO) regions**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>3(0.6%)</td>
<td>2(0.7%)</td>
<td>1(0.5%)</td>
</tr>
<tr>
<td>Western Pacific</td>
<td>142(30.3%)</td>
<td>61(21.4%)</td>
<td>81(43.5%)</td>
</tr>
<tr>
<td>Eastern Mediterranean</td>
<td>16(3.4%)</td>
<td>5(1.8%)</td>
<td>11(5.9%)</td>
</tr>
<tr>
<td>Europe</td>
<td>59(12.6%)</td>
<td>42(14.7%)</td>
<td>17(9.1%)</td>
</tr>
<tr>
<td>South-East Asia</td>
<td>21(4.5%)</td>
<td>10(3.5%)</td>
<td>11(6.5%)</td>
</tr>
<tr>
<td>Americas Region</td>
<td>199(42.4%)</td>
<td>146(51.6%)</td>
<td>52(28.5%)</td>
</tr>
<tr>
<td>Average satisfaction with nursing EHR**</td>
<td>Mean=4.5 (SD=2.3)</td>
<td>Mean=3.7 (SD=2.3)</td>
<td>Mean= 5.6 (SD=2.2)</td>
</tr>
</tbody>
</table>

* Indicates p-value levels <.05 in bivariate comparisons (chi-square/exact Fisher test or t-test) of characteristics between participants who commented on the issues with nursing electronic health records vs. those who did not comment.

** Indicates p-value levels <.001 in bivariate comparisons (chi-square/exact Fisher test or t-test) of characteristics between participants who commented on the issues with nursing electronic health records vs. those who did not comment.

The respondents’ satisfaction with the current state of EHRs

The average satisfaction with the current state of EHRs used by nurses was 4.5 (SD= 2.3) on a scale from 1 (not at all satisfied) to 10 (very satisfied). Some geographic regions (and countries) were represented by a small number of respondents (e.g. from the African countries or Eastern Mediterranean region). We also conducted an association analysis between the respondents’ levels of EHR satisfaction and background characteristics (e.g., level of education, academic & clinical positions, etc.) but did not identify significant associations, thus, these results are not presented here.

The reported issues with the current state of EHRs used by nurses

Two out of three study participants (n= 283, response rate 60.3%) answered question two and provided comments regarding issues with the current state of EHR used by nurses. Table 1 presents a comparison between participants who provided...
comments versus those who did not. Participants who responded to question two had significantly (p<.001, t = 9.1, df = 467) lower scores (mean=3.7, SD= 2.3) for satisfaction with the current state of EHR used for nursing than non-respondents (mean=5.6, SD= 2.2). Overall, participants who commented on the issues with EHR had significantly higher levels of education (e.g., 30.9% of those who commented had PhD vs. 12.4% of non-respondents); had more years of informatics experience; were more likely to be formally trained in informatics; had higher levels of either clinical or academic positions; and were more likely be from the Europe or the Americas Would Health Organizations (WHO) regions, and not the Western Pacific region.

Nine themes were identified in the answers to question two, and each response was mapped to an average of two themes. Each theme was mapped to one of the SV-HIT levels. Overall, 562 concerns were identified in the 283 comments. The most commonly reported concerns were associated with system issues (54.5%), followed by user-task issues (27.5%) and environment issues (18%). Figure 2 presents the distribution of the concerns by SV-HIT levels and Table 2 presents the summary of concern levels, themes and examples of participant quotes.

### User-task issues

**Systems fail to meet nursing clinical needs**: Almost one-third of the respondents (28.6%) indicated that the systems did not meet their clinical needs. One common concern was the inability of the information systems to capture the patient story in either a narrative or structured format. An example comment was, “Too cookie cutter, does not allow for a narrative format, and does not capture the patient story [U.S.]”. Several respondents suggested that further work is needed to improve the utility of the EHR for clinical nursing. For example, “We need to develop better tools that are patient-centered and tell the patient story in ways that are easy and intuitive for healthcare providers to use [U.S.]”.

**Systems are not nursing specific**: Several respondents (16.3%) suggested that information systems do not work well for capturing, storing, and presenting nursing knowledge. Respondents indicated that some of the electronic systems were developed for billing or regulatory reporting needs, thus their usefulness for nursing is limited. For example, respondents suggested that, “Nursing documentation is driven by accreditation & regulatory needs rather than by what actually makes a difference to the patients/families/populations [U.S.]” and “The development of nursing information system is not based on nursing needs, but for the management of hospital expenses [China]”.

Overall, many respondents felt that the current information systems are not capable of reflecting and supporting key aspects of nursing practice. For example, “Systems do not adequately reflect nursing practice, especially the critical thinking and the processing of data and information to knowledge [U.S.]”. In addition, respondents reported that nursing clinical decision support tools were very uncommon. The respondents indicated that, “Current system does not include decision support that is valuable to nurses [U.S.]”.

### System issues

**Poor system usability**: Usability is defined as how easy it is for users to accurately and efficiently accomplish a task while using a system. In our survey, system usability was the most reported concern with almost one-third of the respondents (31.1%) identifying multiple usability issues. The existing systems were referred to as time-consuming and slow, requiring

---

**Figure 2**: Distribution of concerns reported about nursing electronic health records by level*

* Percentage of issues reflect the individual concern level out of total concerns reported, e.g., 127 user-task issues/ 562 total concerns = 27.5%.
too many key strokes to record simple information, and presenting multiple unnecessary screens that interrupt clinical thinking. Multiple respondents experienced issues with poor interface design and challenges with identifying and working with interdisciplinary documentation. Other respondents expressed that documentation is a burden and is often duplicative and hard to track chronologically. Some of the responses included, “The system is not friendly, need to open multiple screens to record information, very slow, and impractical [Brazil]” and “Time consuming, duplicate documentation, lack of discrete data, poor design, too many keystrokes, and nuisance alerts [U.S.]”.

Non-integrated systems and poor interoperability: About a quarter (23.7%) of the respondents indicated that either the systems, or the data that is collected in them, is not integrated with clinical workflows. Some of the common concerns included issues with multiple systems that were difficult to integrate. For example, “Many ‘silos’ and standalone solutions [Sweden]”, “Absence of integrated clinical database [Argentina]” and “Information is spread across many different systems [Belgium]”. One respondent noted that because they use multiple systems in their clinical practice, “Same information is documented in different areas so one ends up double, triple, and even quadruple documenting the same information [U.S.]”.

Lack of interoperability was an additional major concern for survey respondents. Many participants indicated that they require more connected systems to be able to see the patient data across care continuum and collaborate with other professions. For example, “The current systems are not sharing the pertinent information across care areas, such as homecare and acute care, there is also no sharing of information across provinces [Canada]”. Respondents felt that lack of information sharing prevents them from achieving optimal outcomes for their patients. For example, “For care coordination and achievement of the Triple Aim, need more health information exchanges between community - acute - primary care [U.S.]”.

Lack of standards & standardization: Another area of concern reported by some respondents (13.8%) was the lack of use of documentation standards and insufficient system standardization. Several respondents wished to see nursing terminologies implemented to standardize documentation and care quality assessments. For example, “Little comparison of the effect of clinical practice on patient outcomes - no clear way to compare since there is no standardization of terms, [Philippines]” and “Nursing needs to go further and use standardized languages to measure the effectiveness of interventions [Brazil]”. Several nursing terminologies were suggested to resolve the existing issues, such as the International Classification for Nursing Practice (ICNP®). Also, there were a few suggestions regarding the development of guidelines to promote interoperability and system standardization.

Limited functionality/missing components: About one-fifth of the respondents (20.5%) reported that their EHRs lacked at least one key functionality. For example, some suggested that there is not enough coverage for specific content areas, such as pediatric nursing, homecare, or care management. Some other missing areas, mentioned earlier in the paper, included a lack of nursing clinical decision support and inability to re-use nursing data that was already collected for epidemiology and other applied clinical or research processes.

Environment issues

Low prevalence of EHR systems: One-tenth of the respondents (10.6%) indicated that their environments (countries, or a specific health setting) did not implement EHRs comprehensively. For example, “Unfortunately 95% of nursing records are manual records [Venezuela]”; “Few electronic records are being used [South Africa]”; and “Still using paper based systems for nursing documentation with minimal electronic records [Australia]”. Other respondents indicated that their settings, such as homecare, do not have EHRs. A few respondents commented on the differences between regions within countries. For example, “Nursing EHR are only utilized in tertiary private hospitals. Public hospitals in most of the country, especially in provinces, still use paper documentation [Philippines]”.

Lack of users training: Some respondents (7.4 %) identified that insufficient user training prevents full use of EHR system capabilities. A few respondents believed that the generation gap between younger and older nurses requires different levels of training.

Other: Less frequent issues raised by the respondents were grouped as Other. One of the prevalent themes in this category was the lack of EHR support at organizational or policy level. For example, “Policy makers don’t even realize EHR importance so no one is taking any initiatives in this regard [Pakistan]” and “Nursing Councils should motivate Indian Nurses in using EHR [India]”. Others observed a lack of centralized initiatives regarding EHR adoption and/or development. For example, “No national policy to encourage hospitals to develop electronic nursing records [Taiwan]”.

2021
Discussion

Respondents assigned relatively low rankings regarding their satisfaction with the current state of EHRs used by nurses. Two thirds of the respondents specified their concerns with the current systems in their responses to the open ended question. Participants who provided comments to question number two, had lower EHR satisfaction scores, higher levels of education, more years of informatics experience, had higher level professional positions (in clinical practice and academia) and were more likely be from Europe or the Americas WHO regions than participants who did not provide comments. It is possible that more experienced and more educated participants were more likely to respond since they have experienced more EHR issues during their careers.

The literature review identified only a few studies that focused on evaluation of nurses’ use of HIT, and these studies tended to focus on one particular system domain such as usability21-22. Other interdisciplinary studies, mostly focused on medicine, identified similar trends with user perceptions of EHRs21-22. This study revealed low user satisfaction with EHRs that nurses use. More than half of the concerns were at the system issues level of the SV-HIT model, followed by user-task issues and environment issues. These results warrant further examination and potentially, are a call for the attention of international stakeholders (educators, managers, policy makers) to begin to improve the current issues with EHRs used by nurses. In practice, addressing system issues will rely on organizational supports to include end-users in the process of design, purchase, upgrade, and implementation decisions.

Specific user-task issues were identified that affected the ability of information system to support the nurse’s work. First, respondents reported that systems fail to meet nursing clinical needs, such as telling the patient story. The lack of EHRs' ability to tell the patient story is an emerging concern in the healthcare literature regarding system usability23. Respondents also suggested that neither strict structured formats (e.g., check-lists) nor free text documentation (e.g., narrative descriptions) help with clinical decision making. Rather, they hinder a comprehensive understanding of the impact of nursing on patient outcomes. This finding reflects a long-standing discussion regarding how much of the documentation in EHRs should be structured when compared to free text24. One possible solution to this is using automated extraction of important data from free text using natural language processing25. More research on the best balance between narrative and structured documentation to meet nursing clinical needs is needed. Another concerning theme was that EHRs are not nursing specific. Respondents felt that systems were developed either for other disciplines (e.g., medicine) or to meet billing and/or regulatory requirements. Although some literature exists on creating nursing-specific EHRs24, more research and practical recommendations are critically needed.

System issues were the most reported concern about the current EHR systems. In congruence with other studies on the topic21, 22, 27, 28, survey respondents believed that EHR systems suffer from poor usability, e.g., systems are time-consuming, slow, require too many key-strokes and overall burdensome. The survey results also confirmed the well-documented lack of systems interoperability29 and the absence of adoption of nursing terminologies30. Systems lacked key functionality for nursing, with the largest concern being nursing clinical decision support systems. Although some examples of successfully developed and implemented nursing clinical decision support systems exist31, our findings call for more development and application of these systems in clinical practice.

Respondents, mostly from developing countries, indicated that they only have few EHR systems in place. As the trend of EHR adoption increases internationally, driven by local legislations and other reasons, we expect this to change sometime in the next decade. Other environmental level issues included insufficient user training and a lack of EHR implementation support by management and/or policy makers. These issues are also supported by other U.S. and international studies30. In general, there is a need to educate nursing students, practicing nurses, faculty, nurse executives, and the inter-professional care disciplines on key aspects of nursing informatics. There are some existing approaches to increase education and awareness that can help address this issue32-33.

Limitations: Our study has several limitations. First, the generalizability of our survey results is limited due to small number of respondents from certain countries/geographic regions (e.g. from the African countries or Eastern Mediterranean region) and an overrepresentation of nurses with higher professional positions/academic degrees who answered the open ended question (question two). The snowball sampling approach was also limited by the reach of our respective networks and only reached certain organizations and practitioners while others were not included. Our respondents were informaticians and thus were more likely to identify system concerns compared to nurses without an informatics specialization. Further, only a fraction of the survey respondents identified specific issues, such as a lack of documentation standards (13.8%); it is possible that the remainder of participants may have been satisfied with the state of standard terminologies in their EHRs. Finally, respondents who answered the open ended question had a lower mean EHR satisfaction score and were mostly from Europe and North and South America, with fewer responses from Asia and Pacific.
Learning from these limitations, we are currently planning a follow-up study to explore EHR-related international concerns in more detail.

Conclusions

This study focused on nurses’ satisfaction and issues with EHRs. It is one of the largest studies of international trends in nursing informatics. Respondents from 45 countries ranked their satisfaction with the current state of EHRs relatively low. Two-thirds of more educated and more experienced study participants, mostly from the Europe or the Americas WHO regions, provided disconcerting comments explaining their low EHR rankings. More than one-half of the comments identified issues at the system level (e.g., poor system usability; lack of integrated systems and poor interoperability; lack of standards & standardization; and limited functionality/missing components), followed by user–task issues (e.g., systems fail to meet nursing clinical needs; and systems are not nursing specific) and environment issues (e.g., low prevalence of EHR systems; and lack of user's training). Although the study sampling and analytical approaches have limitations, the results call for the attention of international stakeholders (educators, managers, policy makers) to begin to improve the current issues with EHRs used by nurses.

References


2023
<table>
<thead>
<tr>
<th>Level</th>
<th>Theme</th>
<th>Number of comments (% out of total comments)*</th>
<th>Example quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER-TASK ISSUES</td>
<td>Systems fail to meet nursing clinical needs</td>
<td>81 (28.6)</td>
<td>&quot;The development of nursing information system is not based on nursing needs [China]&quot;; &quot;Hard to tell the patient story [U.S.]&quot;; &quot;Continuing invisibility of nursing's contribution to outcomes [U.S.]&quot;; &quot;Currently doesn't feel like it enhances live patient care [Australia].&quot;</td>
</tr>
<tr>
<td></td>
<td>Systems are not nursing specific</td>
<td>46 (16.3)</td>
<td>&quot;Nursing documentation is driven by accreditation &amp; regulatory needs rather than by what actually makes a difference to the patients/families/populations [U.S.]&quot;; &quot;[Systems] Need more focus on nursing care (processes, measurable tasks, outcomes associated with nursing Interventions) - needs to be more intuitive to nursing knowledge [U.S.]&quot;; &quot;Systems do not adequately reflect nursing practice, especially the critical thinking and the processing of data and information to knowledge [U.S.].&quot;</td>
</tr>
<tr>
<td>SYSTEM ISSUES</td>
<td>Poor system usability</td>
<td>88 (31.1)</td>
<td>&quot;Poor interface design [Taiwan]&quot;; &quot;Very time consuming and duplicative[Finland]&quot;; &quot;Documentation can be hard to find. Especially in non discrete data driven areas of the hospital- like behavioral health, social services, physical and speech therapies [U.S.]&quot;; &quot;Too many keystrokes, and nuisance alerts [U.S.].&quot;</td>
</tr>
<tr>
<td></td>
<td>Non- integrated systems and poor interoperability</td>
<td>67 (23.7)</td>
<td>&quot;[Systems are] Scattered, hard to integrate [China]&quot;; &quot;For care coordination and achievement of the Triple Aim, Need more Health Information Exchanges between community - acute - primary care [U.S.]&quot;; &quot;Lack of interoperability with other systems [U.S.].&quot;</td>
</tr>
<tr>
<td></td>
<td>Lack of standards &amp; standardization</td>
<td>39 (13.8)</td>
<td>&quot;There is very little to no standardization on documentation to facilitate information management and retrieval [Philippines]&quot;; &quot;Little, if any, implementation or use of standardized terms outside of physician practice (CPT or ICD-9/ICD-10) [U.S.]&quot;; &quot;Little comparison of the effect of clinical practice on patient outcomes - no clear way to compare since there is no standardization of terms[Philippines]&quot;; &quot;A great deal of information is not standardized without standard terminologies [U.S.].&quot;</td>
</tr>
<tr>
<td></td>
<td>Limited functionality/missing components</td>
<td>58 (20.5)</td>
<td>&quot;Missing process support and a long list of small details, missing functionality for check lists [Sweden]&quot;; &quot;Care Management options are limited [U.S.]&quot;; &quot;Pediatric content is limited in both Home Health and Hospital systems and must be created [U.S.].&quot;</td>
</tr>
<tr>
<td>ENVIROMENT ISSUES</td>
<td>Low prevalence of EHR systems</td>
<td>30 (10.6)</td>
<td>&quot;Unfortunately 95% nursing records are manual records [Venezuela]&quot;; &quot;Few electronic records are being used [South Africa]&quot;; &quot;Still using paper based systems for nursing documentation with minimal electronic records [Australia].&quot;</td>
</tr>
<tr>
<td></td>
<td>Lack of user's training</td>
<td>21 (7.4)</td>
<td>&quot;Users don't know how to use the system [U.S.]&quot;; &quot;Need more awareness and inservice training for Indian Nurses [India]&quot;; &quot;Only very few professionals are qualified or trained to practice computerized documentation [Philippines].&quot;</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>32 (11.3)</td>
<td>&quot;Lack of participation from clinical staffs in system development [Taiwan]&quot;; &quot;I know nothing about how nursing EHR is created because it is controlled by our vendor [U.S.]&quot;; &quot;No support from the authorities [Argentina]&quot;; &quot;Poor understanding of the implementation of computerized documentation in nursing, both by managers as the professionals themselves [Brazil].&quot;</td>
</tr>
<tr>
<td><strong>Total comments</strong></td>
<td></td>
<td>283</td>
<td></td>
</tr>
</tbody>
</table>
Why aren’t they happy? An analysis of end-user satisfaction with Electronic health records

Prasad Unni1, MD, Catherine Staes1, BSN, MPH, PhD, Howard Weeks1, MD, Heidi Kramer1, PhD, Damion Borbolla1, MD, MS, Stacey Slager1, MS, Teresa Taft1, BA, Valliammai Chidambaram1, MS, Charlene Weir1,2, PhD, RN
1University of Utah, School of Medicine, Salt Lake City, Utah
2SLC VA IDEAS Center of Innovation, Salt Lake City, Utah

Abstract

Introduction: Implementations of electronic health records (EHR) have been met with mixed outcome reviews. Complaints about these systems have led to many attempts to have useful measures of end-user satisfaction. However, most user satisfaction assessments do not focus on high-level reasoning, despite the complaints of many physicians. Our study attempts to identify some of these determinants.

Method: We developed a user satisfaction survey instrument, based on pre-identified and important clinical and non-clinical clinician tasks. We surveyed a sample of in-patient physicians and focused on using exploratory factor analyses to identify underlying high-level cognitive tasks. We used the results to create unique, orthogonal variables representative of latent structure predictive of user satisfaction.

Results: Our findings identified 3 latent high-level tasks that were associated with end-user satisfaction: a) High-level clinical reasoning b) Communicate/coordinate care and c) Follow the rules/compliance.

Conclusion: We were able to successfully identify latent variables associated with satisfaction. Identification of communicability and high-level clinical reasoning as important factors determining user satisfaction can lead to development and design of more usable electronic health records with higher user satisfaction.

Introduction

Healthcare institutions have been increasingly implementing electronic health records (EHR) over the last decade. These systems are very expensive and have mixed reviews. Despite many positive outcomes, clinicians continue to complain about difficulties with usability, lack of cognitive support or failures to match workflow in design. EHRs that provide cognitive support should be expected to match the users’ task-based mental processes such as perception, memory, reasoning and behaviour.

The degree to which systems fail to support clinical reasoning, interrupt complex thought processes and increase (rather than decrease) cognitive load have been cited as a major source of dissatisfaction. Studies show that irrelevant display clutter can negatively affect the clinicians’ ability to perform tasks and lead to clinicians missing critical information. Studies show that irrelevant display clutter can negatively affect the clinicians’ ability to perform tasks and lead to clinicians missing critical information.

User satisfaction is a combination of the ease of use and the degree to which the system supports work and is useful. In this paper, we are focusing on how well the system supports user’s perceived performance of high-level reasoning tasks. We created and validated an instrument based on high-level reasoning tasks as one way to explore underlying mental representations.

Background

Dissatisfaction with EHR Design

EHR systems with poor designs significantly increase the mental workload of a clinician while performing high-level cognitive tasks thereby reducing user satisfaction, increasing frustration, and causing ineffective work-arounds. Such poorly designed systems can also affect patient safety. A 2009 report by National Research Council of deployed EHRs studied how health information technology (HIT) is used in 8 different medical centers and noted that none provided the much needed cognitive support to clinicians, i.e. the high-level reasoning and decision.
making capacity that spans several low-level transactional tasks like ordering and prescribing. As per the report, IT applications are largely designed to automate simple clinical actions or business processes. These narrow task-specific automation systems that make up the majority of today’s health care IT provide little cognitive support to clinicians. For example, as Stead and Linn noted:

“Today, clinicians spend a great deal of time and energy searching and sifting through raw data about patients and trying to integrate these data with their general medical knowledge to form relevant mental abstractions and associations relevant to the patient’s situation. Such sifting effort forces clinicians to devote precious cognitive resources to the details of data and make it more likely that they will overlook some important higher-level consideration.”

This refrain is common amongst clinicians, but the exact nature of these higher-level “considerations” is not clear. Although EHR adoption has been steadily increasing, in part due to federal incentives, research suggests that the design and implementation of EHRs do not align with the cognitive and/or workflow requirements and preferences of physicians.

Prior research on identifying the cognitive tasks involved in healthcare delivery has come from a variety of traditions. Ethnographic studies assessing user’s needs in a clinical environment lists behavior actions, but often they do not make sense clinically and cognitively. For example, Tang, Jaworski and Fellencer, et al. (1996) observed 38 clinicians across several sites and found 5 types of activities: reading, writing, talking, and assessing the patient. Although accurate, these kinds of descriptions of activity types do not translate to understanding cognitive processing goals.

Complexity in health-care decision-making arises out of a multitude of factors, including sicker patients, general increase in the prevalence of chronic disease, and the increased emphasis on team care. The delivery of high-quality health care to patients increasingly involves multiple clinicians – primary care physicians, specialists, nurses, technicians and others – all of whom are required to deliver effective coordinated care. Each clinician has specific, limited interactions with the patient and thus a somewhat different view resulting in increased fragmentation or compartmentalized into silos of facts and disconnected symptoms. Although ONC (Office of the National Coordinator for Health IT) encourages the use of EHRs across the continuum of care, problems in continuity of care continue.

A qualitative study conducted by AHRQ on clinician’s recommendations for improving clinical communication and patient safety, found that HIT and organization of EHR information were the most frequent safety solutions recommended by clinicians to address problems in clinician-to-clinician communication. Moreover, clinicians envisioned EHRs to become the desired “one source of truth” for a patient’s medical information. How this source of truth is organized, accessed and displayed remains problematic.

**Mental Representations of Tasks**

Authors in information systems research examining "task-technology fit" (TTF) also focus on task definition. TTF was proposed early by Goodhue and others and expanded by Ammenworth and Lee with the goal of examining the degree to which a system helps to support user’s tasks. The premise is that satisfaction (and use) is derived from a good match between the tasks and the technology. Some of this work examined the impact of data representation on task performance and found performance dropped rapidly when data representation did not reflect cognitive processing. Some of this work focused on dimensions of tasks and not the cognitive representation of behavior.

In this paper, we approach defining clinical tasks based on the Theory of Action Identification. Assessing the clinician’s high-level cognitive actions or tasks require exploring the mental representation of action. According to Action Identification Theory, behavior is controlled by hierarchical representations that are a group of associated neurons and pathways that include behaviors, schemas, and goals. Specific movements involved in an action, such as typing or searching or scrolling are at the “bottom.” The higher-level component of the representation includes the goals, intentions and associated values. A core principle is that expert behavior is controlled and cued at the highest levels possible. In other words, when physicians start to investigate the best treatment for a cancer or they initiate a plan for helping the patient control their diabetes, they are not really engaged in typing, sifting, sorting and scrolling or even in looking for Hemoglobin A1C trends over time. They are actually looking to see if the treatment protocols are working. That high-level action identification is supported by a myriad of smaller information representations and physical actions. The other key aspect of this theory is that behavior that is functioning at the highest level of its representation is more fluid, uses less cognitive resources and is more satisfying. Think about what happens to an expert musician when you ask them to think about where to put their fingers. They may become
distracted and lose focus. Similarly, high-level reasoning actions are also interrupted when they cannot “flow” due to an increase in the difficulty of actually performing the task.

The high-level identities of many clinical tasks have not been made explicit. As a result, designing EHR to support those tasks is difficult. In addition, evaluating the success of systems to support those tasks remains indirect and is often limited to simple usability testing. General usability metrics often neglect high-level reasoning tasks or activities and do not discriminate between high-level reasoning difficulties and those difficulties associated with the interface itself.

**Objectives**

In this study, our goal was to explore the characteristics of high-level tasks by generating task-based questions for a survey from prior work (observations, interviews, and surveys) and validating the instrument using factor analysis to determine latent task structures. If the factor analysis revealed a clustering of items together that reflected the tasks, then we would have established one form of validation. We used exploratory factor analysis, an analytic technique used for examining latent structures in a set of items. The ultimate goal is to develop a method of measuring user satisfaction that addresses high-level reasoning tasks and provides useful directions for design of an effective EHR.

**Method**

**Design:** The study was conducted in 3 phases. First, we identified an initial set of 51 questions addressing clinician tasks from prior literature and our own work. The top most frequent and important questions were preserved after pilot testing and were used to build the final survey instrument. The final survey was conducted among practicing clinicians across all levels of service in multiple specialities. Second, an exploratory factor analysis of the survey results was conducted. Finally, scales were created using tests of Cronbach’s alpha reliability and evaluation of item loadings on derived factors. The aggregate ratings of those scales were used to predict the overall judgment of how well the system supported work performance. Specific details are outlined in results.

**Setting and Sample**

The University of Utah Institutional Review Board approved this study. Data collection was conducted at a large, >500-bed academic hospital located in Salt Lake City, Utah. We identified and targeted 189 residents, fellows and attendings who were currently posted at the University of Utah hospitals or the Huntsman Cancer Institute and were using EPIC EHR for their in-patient care. We used direct approach, email public link to the survey or email personal link to contact the targeted participants for the survey. All email links were created using the REDCap survey tool.

Residents were the targeted participants were approached during one month, in December 2015, 18 months following EPIC implementation.

For purposes of statistical analysis service domains were broadly categorized into “Medical Specialties” which includes: Internal Medicine, Anesthesiology, Physical Medicine and Rehab, Psychiatry, Neurology, Pediatrics, Cardiology, Gastroenterology, Rheumatology and Family Medicine, and “Surgical Specialties” which includes: General Surgery, Cardiovascular Surgery, Obstetrics & Gynecology, Neuro Surgery, Plastic Surgery, Orthopedics and Urology.

**Phase I: Item Selection and Evaluation**

An initial set of 51 tasks were created based on our prior research and information needs as well as from the general recommended areas of concern noted by Stead and Linn. This report highlighted 7 information sensitive domains identified in Institute of Medicine’s vision for 21st century healthcare: a) integrate patient information from various sources, b) integrate evidence base with daily practice, c) developing tools for portfolio management of patients at an individual and population level, d) rapid integration of new knowledge into a ‘learning’ health system, e) extending care (treatment and monitoring) beyond hospitals through better communication and technology, thereby empowering patients in the decision making process more effectively. The items were all identified as tasks that clinicians needed to accomplish during the course of their work. We designed the questions to be about at the same level of abstraction and to reflect general domains of work, e.g. patient assessment, diagnosis, teamwork.

The items were extensively pilot-tested and then rated by 8 physicians as to their frequency/importance on a 1 (Very Low) to 5 (Very High) scale. The final set of questions included those that were rated in the top 50%. (n=21).
The final survey questionnaire had 3 parts. The first part included demographic questions about the participant including name, department affiliation and level year. For the second part, clinicians were asked to respond to the question “While caring for hospitalized patients, how well does Epic help you to…” The responses were captured in a Likert scale ranging from 1 (very poorly) to 5 (very well). The third part consisted of a free text comment box, where the user was asked to add any additional comments he/she had. The questions used in the survey instrument are listed in Table 2.

Phase II: Validation of Factor Structure and Variable Creation

Factor analysis has long been used in the social sciences to assess for underlying latent structure in perceptions and mental variables which cannot be measured directly (e.g., intelligence). Multiple methods have been devised for determining these factors (latent variables), all of which center on determining the pattern of correlations between variables. A pattern of correlations amongst the questions may reveal clusters that suggest those questions are measuring the same construct. Inspection of the resultant patterns of correlations (factors) provide evidence of a latent structure. We conducted an exploratory factor analysis using a correlation matrix with varimax rotation. Recommendations for the sample size varies with some recommending as low as 5 participants per question and as many as 10 or more. Our analysis showed each common factor was over-determined with at least 5 measured variables representing each common factor. As few as 3-4 measured variables per common factor can reduce the required sample size to approximately 100, which is consistent with our sample size.

Items loading on the same factor with a loading value of greater than 0.5 and not loading greater than 4.0 on another factor were combined (summed and averaged) to create 3 new variables (or scales). The reliability of these new variables were assessed using Cronbach’s alpha with an acceptability criteria of greater than 0.80.

Phase III. Predictive Validity Assessment

The 3 newly defined variables were then regressed on the rating of overall efficiency to validate their association with performance using linear regression.

Results

Study population

Of the 189 targeted clinicians (residents, fellows, and attendings) who were working in the system at the time, 47.6% (n=90) participated in the survey. The distribution of respondents by level and domain are presented in Table 1 below. Two participants did not disclose their service category (department) and one participant did not disclose the level year of service.

Table 1. Participant sample size (n=90) based on broad speciality and level of service

<table>
<thead>
<tr>
<th>Level</th>
<th>Surgery (n)</th>
<th>Medicine (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGY 1</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>PGY 2</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>PGY 3</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>PGY 4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>PGY 5</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>PGY 6</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>PGY 7</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Fellow</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Attending</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>57</td>
</tr>
</tbody>
</table>

PGY – Post Graduate Year

2029
**Descriptive Results**

The descriptive statistics of each survey question (n, range, min, max, mean and SD) are listed in Table 2. The number of respondents to each question ranged from 70 to 90, with a mean satisfaction score ranging from 2.88 to 3.85 (Max 5)

**Table 2 – Survey Questionnaire with descriptive results**

<table>
<thead>
<tr>
<th>Survey questions – “How well does Epic help you to...”</th>
<th>N</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Efficiently do your work overall</td>
<td>90</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.60</td>
<td>.909</td>
</tr>
<tr>
<td>2. Discover why the patient is receiving a specific treatment and/or medication</td>
<td>90</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.13</td>
<td>.997</td>
</tr>
<tr>
<td>3. Identify a patient's current severity of illness</td>
<td>90</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.07</td>
<td>.958</td>
</tr>
<tr>
<td>4. Track clinical parameters that need continuous monitoring</td>
<td>90</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.74</td>
<td>.966</td>
</tr>
<tr>
<td>5. Review appropriateness for drug therapy</td>
<td>88</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.15</td>
<td>.891</td>
</tr>
<tr>
<td>6. Prevent and monitor for adverse drug events</td>
<td>85</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.31</td>
<td>.900</td>
</tr>
<tr>
<td>7. Integrate external sources of patient information</td>
<td>83</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.01</td>
<td>1.110</td>
</tr>
<tr>
<td>8. Comprehend the relationship and dependencies between medication and labs</td>
<td>86</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.10</td>
<td>.921</td>
</tr>
<tr>
<td>9. Follow the sequence and timeline of a clinical event</td>
<td>89</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.33</td>
<td>.963</td>
</tr>
<tr>
<td>10. Integrate patient preferences into the care plan</td>
<td>86</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.06</td>
<td>.998</td>
</tr>
<tr>
<td>11. Document my plan of care</td>
<td>88</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.85</td>
<td>.865</td>
</tr>
<tr>
<td>12. Communicate goals of care to the team</td>
<td>89</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.49</td>
<td>.919</td>
</tr>
<tr>
<td>13. Determine if the patient's team are on the same page and agree on the clinical plan</td>
<td>88</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.02</td>
<td>1.005</td>
</tr>
<tr>
<td>14. Communicate results of hospitalization to the Primary Care Provider</td>
<td>83</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.27</td>
<td>1.049</td>
</tr>
<tr>
<td>15. Organize information to order a referral</td>
<td>83</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.30</td>
<td>.959</td>
</tr>
<tr>
<td>16. Write an effective discharge summary</td>
<td>80</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.55</td>
<td>.940</td>
</tr>
<tr>
<td>17. Comp with billing requirements</td>
<td>70</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.63</td>
<td>.871</td>
</tr>
<tr>
<td>18. Track my compliance with guidelines in treating current patients</td>
<td>75</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2.97</td>
<td>.944</td>
</tr>
<tr>
<td>19. Identify rules for admission/discharge/transfer</td>
<td>77</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2.88</td>
<td>.986</td>
</tr>
<tr>
<td>20. Identify patients who require a higher acuity level of care</td>
<td>83</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.17</td>
<td>.922</td>
</tr>
<tr>
<td>21. Access relevant clinical knowledge at Point of care</td>
<td>87</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3.13</td>
<td>1.043</td>
</tr>
</tbody>
</table>

**Level of service**

Participant sample for this study included clinicians from post graduate year 1 to post graduate year 7, fellows and attending physicians. Total user satisfaction score was calculated for all the participants and were categorized based on the year of service. Average satisfaction score for each level of service was calculated by dividing the total score of all participants in each level year by the product of total number of participants in each level year and the number of questions answered (Figure 1).
Factor Analysis

An exploratory factor analysis using the correlation matrix was conducted using varimax rotation. The number of factors was constrained to 3 based on inspection of the scree plot showing 1 factor with a large eigenvalue (amount of variance) and 2 smaller factors (which we preserved for theoretical reasons). There were 4 factors with eigenvalues above 1.0; however the lowest one explained only 5.26% of the variance with an eigenvalue of only 1.04 so it was excluded. Eigenvalues are a measure of how much of the observed variables’ variance is explained by the factor. The initial eigenvalues show that the first factor explained 48.85% of the variance, the second factor 6.75% of the variance, the third factor 6.11% of the variance and the fourth factor 5.26% of the variance (see Table 3). Any factor with an eigenvalue ≥ 1 explains more variance than a single observed variable and is the minimal acceptable level to be included as a factor. The first 3 factors explained 62% of the variance.

Table 3. Factors, eigenvalues and variance

<table>
<thead>
<tr>
<th>Factor</th>
<th>Initial Eigenvalues</th>
<th>% of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.26</td>
<td>48.85</td>
</tr>
<tr>
<td>2</td>
<td>1.42</td>
<td>6.75</td>
</tr>
<tr>
<td>3</td>
<td>1.22</td>
<td>6.11</td>
</tr>
<tr>
<td>4</td>
<td>1.04</td>
<td>5.26</td>
</tr>
</tbody>
</table>

Table 4 displays the items and their associated factor loadings. The questions related to each of the 3 factors were examined and the factors were identified as: 1) High-level clinical reasoning (i.e.: complex thinking about patient information, such as linking meds and labs); 2) Communication /coordination (i.e.: those items associated with shared responsibility of care, including communicating with team or the primary care physician (PCP)); and 3) Following the rules / compliance (i.e.: following billing rules/ following clinical guidelines). These labels are descriptive of the underlying construct. Two questions did not show a strong association (above 0.5) with any factors (questions 6 and 7).

Scale Construction

Composite scores were created for each of the 3 factors by summing the items highlighted in Table 4. Higher scores indicate higher satisfaction. Reliability was tested for each item. All were acceptable as measured by Cronbach’s alpha (see Table 5).
Table 4. Factor loadings and communalities for survey questions

<table>
<thead>
<tr>
<th>Qn no</th>
<th>Survey Questions</th>
<th>Factor 1 – Communication/Coordination</th>
<th>Factor 2 – High-Level Clinical Reasoning</th>
<th>Factor 3 – Rule/Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Care plan documentation</td>
<td>0.807</td>
<td>0.026</td>
<td>0.322</td>
</tr>
<tr>
<td>14</td>
<td>Communicate with PCP</td>
<td>0.703</td>
<td>0.304</td>
<td>-0.079</td>
</tr>
<tr>
<td>13</td>
<td>Patient’s team sync</td>
<td>0.687</td>
<td>0.209</td>
<td>0.440</td>
</tr>
<tr>
<td>12</td>
<td>Communicate goals</td>
<td>0.671</td>
<td>0.321</td>
<td>0.323</td>
</tr>
<tr>
<td>9</td>
<td>Follow timelines</td>
<td>0.617</td>
<td>0.408</td>
<td>0.336</td>
</tr>
<tr>
<td>4</td>
<td>Track Parameters</td>
<td>0.565</td>
<td>0.013</td>
<td>0.378</td>
</tr>
<tr>
<td>1</td>
<td>Work Efficiency</td>
<td>0.537</td>
<td>0.286</td>
<td>0.224</td>
</tr>
<tr>
<td>15</td>
<td>Info for Referral</td>
<td>0.515</td>
<td>0.432</td>
<td>0.452</td>
</tr>
<tr>
<td>7</td>
<td>Integrate external</td>
<td>0.499</td>
<td>0.461</td>
<td>0.077</td>
</tr>
<tr>
<td>8</td>
<td>Med-Lab relation</td>
<td>0.076</td>
<td>0.758</td>
<td>0.107</td>
</tr>
<tr>
<td>10</td>
<td>Patient Preference</td>
<td>0.239</td>
<td>0.676</td>
<td>0.312</td>
</tr>
<tr>
<td>20</td>
<td>Identify Acuity need</td>
<td>0.321</td>
<td>0.655</td>
<td>0.396</td>
</tr>
<tr>
<td>3</td>
<td>Identify severity</td>
<td>0.459</td>
<td>0.651</td>
<td>0.145</td>
</tr>
<tr>
<td>5</td>
<td>Drug Appropriateness</td>
<td>0.091</td>
<td>0.620</td>
<td>0.473</td>
</tr>
<tr>
<td>2</td>
<td>Why Treatment</td>
<td>0.506</td>
<td>0.592</td>
<td>0.128</td>
</tr>
<tr>
<td>6</td>
<td>Adverse Drug Events</td>
<td>0.439</td>
<td>0.460</td>
<td>0.338</td>
</tr>
<tr>
<td>18</td>
<td>Guideline Compliance</td>
<td>0.261</td>
<td>0.365</td>
<td>0.773</td>
</tr>
<tr>
<td>17</td>
<td>Billing Compliance</td>
<td>0.166</td>
<td>0.028</td>
<td>0.731</td>
</tr>
<tr>
<td>19</td>
<td>Identify ADT rules</td>
<td>0.146</td>
<td>0.477</td>
<td>0.681</td>
</tr>
<tr>
<td>21</td>
<td>Knowledge at POC</td>
<td>0.321</td>
<td>0.362</td>
<td>0.622</td>
</tr>
<tr>
<td>16</td>
<td>Discharge Summary</td>
<td>0.439</td>
<td>0.248</td>
<td>0.569</td>
</tr>
</tbody>
</table>

Table 5. Descriptive statistics of aggregated variables

<table>
<thead>
<tr>
<th>Factors</th>
<th>No. of items</th>
<th>Mean</th>
<th>SD</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication/Coordination</td>
<td>8</td>
<td>3.42</td>
<td>0.76</td>
<td>0.861</td>
</tr>
<tr>
<td>High-level clinical reasoning</td>
<td>6</td>
<td>3.16</td>
<td>0.73</td>
<td>0.875</td>
</tr>
<tr>
<td>Compliance/Follow the Rules</td>
<td>5</td>
<td>3.23</td>
<td>0.74</td>
<td>0.850</td>
</tr>
</tbody>
</table>

Regression and Predictive Validity

The 3 constructed variables, communication/coordination, compliance/follow the rules and high-level clinical reasoning were regressed on the rating of overall efficiency using forward entry. The overall model was significant (R=0.54; p=0.00) with communication entering in first with the largest component and high-level clinical reasoning entered on the second step producing a significant R2 change (R=0.59; p=0.04). Compliance (follow the rules) was not significant after the other 2 variables were included in the model.
**Discussion**

Our exploratory factor analysis successfully explained 62% of the overall variance in the items and 3 latent high-level cognitive tasks were validated with the factor structure. These can be thought of as latent variables referring to overarching “groups” of actions. They also illustrate how clinicians are trying to use the EHR to accomplish their work and identify high-level intentions for action in the clinical environment.

The overall means across these 3 variables were not meaningfully different, although they correlated significantly. All were in the middle of the scale (around 3.5) indicating mid-level of satisfaction. The validity of the constructs were also supported by the regression results with communication/coordination and high-level reasoning explaining most of the variance in overall satisfaction.

Current user satisfaction studies focus on specific functionalities and technical attributes of the system rather than a deep understanding of the cognitive processes that influence the end-users’ usage of the system. Specific models like the Technology Acceptance Model (TAM) suggests user perceptions of ease of use and usefulness as major determinants of user satisfaction. Additionally, Delone and McLean information system success model suggests other determinants of user satisfaction like user characteristics, a) system usefulness and b) service quality. None of these address satisfaction regarding the ability of the system to support high-level cognitive work. A popular usability measurement instrument, System Usability Scale (SUS), is a 10 question survey instrument which focuses on the user perception of a system’s usability and learnability metrics. However there is very little information that can be inferred from the SUS regarding the system’s ability to support high-level cognitive tasks, communicability or compliance with regulations.

Our research strongly suggests that user satisfaction is determined by factors well beyond those suggested by the above models. We identified 3 high-level tasks factors: a) ability to communicate/coordinate care between teams, b) ability to conduct cognitive tasks and c) ability to abide by the rules, as independent determinants that significantly influence the end-user satisfaction with our EHR system. Our findings validate the common clinician complaints about the lack of support to carry out cognitive tasks and collaborate between teams observed in previous research. Amongst the 3 high-level task factors, we found that communication/coordination had the strongest relationship with overall efficiency leading to highest perceived user satisfaction.

These findings suggest that the ability of EHRs to support these high-level tasks could be one crucial factor in determining overall user satisfaction. However, we may not have identified all of the attributes related to high-level action identities. Different contexts would likely be associated with more variety in the specific high-level tasks in which clinicians are engaged as well as the amount of non-EHR support. This pattern would explain why a system may be well received by users within one context, but rejected in another. As clinicians’ tasks vary from one setting to another, system vendors and system optimizers should modify their current approach and try to understand this dynamism and optimize the system accordingly.

**Implications for Providing Cognitive Support**

Clinicians’ cognitive resources are limited, and therefore any task which forces them to divert their attention away from patient care to data acquisition and interpretation should be avoided. Functionalities that support coordination, communication, shared views and negotiated responsibilities might be more important than realized. In addition, systems should be able to provide a holistic picture of the data in a complete patient context and its relevance to the situation at hand. Having this holistic picture can significantly reduce the cognitive effort required while conducting high-level tasks.

**Clinical Implications**

Our study shows that ability to communicate and coordinate between various clinical teams with relative ease and efficiency is associated with significantly higher user satisfaction amongst clinicians. As healthcare evolves from an individual doctor – patient relationship to a multidisciplinary team-based approach, the functionality of electronic health records to support collaboration between the teams proves to be crucial in a successful EHR implementation. Another important factor that determines user satisfaction was found to be the ability of the system to support high-level cognitive tasks conducted by the clinicians. These findings should prompt EHR vendors to undertake further system modifications to support these functionalities. It would also encourage clinical informaticians to undertake system optimization projects where they could study the various cognitive tasks within different contexts, and to implement practical solutions.
**Limitations**

The sample size to support a factor analysis was only minimally acceptable. The use of one variable assessing overall satisfaction might limit reliability. Finally, the study was conducted at only one site, which may impact generalizability. Even so, this study was intended to be exploratory and involved a newly-developed instrument which can be used and further validated by others.

**Conclusion**

Our research significantly increases our understanding of the determinants of end-user satisfaction with electronic health record (EHR) systems. We find that measuring a system’s ability to carry out high-level tasks has a much higher influence on end-user satisfaction than currently measured determinants. These findings could be adopted by system vendors as well as system optimizers in the future to create effective and efficient electronic health records that generate true value to the end-user.

**Acknowledgements**

This study was funded by the University of Utah Health Science Center to the Department of Biomedical Informatics.

**References**


2035
Learning Effective Treatment Pathways for Type-2 Diabetes from a clinical data warehouse

Rohit Vashisht, PhD1, Ken Jung, PhD1, Nigam Shah, MBBS, PHD1

1Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA, United States

Abstract

Treatment guidelines for management of type-2 diabetes mellitus (T2DM) are controversial because existing evidence from randomized clinical trials do not address many important clinical questions. Data from Electronic Medical Records (EMRs) has been used to profile first line therapy choices, but this work did not elucidate the factors underlying deviations from current treatment guidelines and the relative efficacy of different treatment options. We have used data from the Stanford Hospital to attempt to address these issues. Clinical features associated with the initial choice of treatment were effectively re-discovered using a machine learning approach. In addition, the efficacies of first and second line treatments were evaluated using Cox proportional hazard models for control of Hemoglobin A1c. Factors such as acute kidney disorder and liver disorder were predictive of first line therapy choices. Sitagliptin was the most effective second-line therapy, and as effective as metformin as a first line therapy.

Keywords: Type-2 Diabetes, Learning Health Systems, Treatment Pathways, Second-line treatment options

Introduction

Type 2 diabetes mellitus (T2DM) affects an estimated 29.1 million people in the United States [1]. Its global prevalence is projected to reach 440 million adults by the end of 2030 [2]. Current treatment guidelines, which are derived from a few randomized controlled trials [3-6], recommend the use of metformin as first line mono-therapy [7]. However, when metformin exhibits adverse effects or fails to control diabetes, the second line therapy must be chosen, and there is little consensus on how to choose a second line therapy; with the American Diabetes Association recommending sulfonylureas, meglitinide, pioglitazone or dipeptidyl peptidase 4 inhibitor (DPP4) as second-line agent [8], and the American Association of Clinical Endocrinologists recommending alpha-glucose inhibitors, DPP4 inhibitors and GLP-1 agonist [9]. Given the availability of myriad treatment options for second-line therapy, the problem of selecting an optimal second-line agent requires urgent attention.

In contrast, knowledge captured during routine clinical care in Electronic Medical Records (EMRs) has ushered in a new era of learning health systems that can provide evidence for medical decision-making beyond that from formal clinical studies [10]. Initiatives such as the Patient-Centered Research Institute (pcornet.org), Observational Health Data Science & Informatics program (OHDSI) (ohdsi.org) and the Green Button [11] exemplify early efforts towards building learning health-care systems. In one recent study, analysis of data by the Observational Health Data Science and Informatics initiative revealed significant diversity in the choice of first line therapy for T2DM [12]. Harnessing data from 11 databases that collate 250 million records into a unified common data model, the study found that metformin was the predominant initial choice of therapy but that other choices were also common. Substantial heterogeneity in the prescription of second-line agents was also noted, highlighting a gap in available clinical guidelines for management of T2DM. Important clinical questions such as the factors that determine the initial choices of treatment, and the effectiveness as well as the optimal choice for second-line therapy were not addressed.

In pursuit of these goals, we set out to perform a systematic analysis of treatment-pathways in T2DM using data collated in Stanford’s clinical data warehouse[13]. We recapitulate previous work regarding variation in the choice of first line therapy, and find clinical factors that are predictive of the first line therapy choice that are consistent with biomedical knowledge of adverse effects associated with metformin. Finally, we
demonstrate the feasibility of assessing the relative efficacy of different choices of second line therapy in controlling HbA1c. We find that sitagliptin appears to be as effective as metformin as a first line therapy, and is considerably better than other options as a second line therapy. At present, our analysis is limited to data from Stanford Hospital, and is an example of a research study that can be extended to any site that has adopted the OHDSI common data model.

Methods
Data Source
The electronic medical records of 2.1 million patients were obtained from the Stanford Clinical Data Warehouse under an approved IRB protocol [13]. Patient satisfying the following creation were selected for our study cohort. The index date of each patient was defined as the date of the first prescription of any of the drugs considered in this study following at least one mention of a T2DM ICD9 code. Patient were required to have at least 90 days of data prior to their index date. Patients were also required to have at least one HbA1c measurement following their index date to be included in the analysis of treatment pathway efficacy.

Practice-Based Treatment Pathways
Our approach extracts sequences of drug prescriptions from the EMR, reflecting treatment decisions made in actual clinical practice. We call such sequences “practice based treatment pathways”, and they are the basis of our analysis of factors influencing the choice of initial treatment as well as the relative effectiveness of different treatment options. First, we construct a cohort of diabetic patients using criteria described above and summarized in Figure 1a. Treatment pathways were constructed by following the medication trajectory of patient from their respective index dates.

Our analysis was focused on two types of treatment pathways. First, in order to focus on the factors influencing the choice of first line therapies and their efficacy, we considered treatment pathways consisting of a single drug. Second, we considered treatment pathways in which the second prescription was different from that of first, with the first prescription lasting at least 30 days. Current guidelines suggest intensifying therapy after two months of the initial therapy if the desired HbA1c goals are not met [9]. At present we restricted our analysis to have required a minimum gap of at least 30 days before the prescription of next drug, which can be relaxed in future analysis. The treatment pathways with a second prescription were used to evaluate the efficacy of second line treatments.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Drug Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metformin</td>
<td>Biguanide</td>
</tr>
<tr>
<td>Glipizide, Glyburide, Tolazamide, Acetohexamide, Glucotrol, Chlorpropamide, Glimepiride</td>
<td>Sulfonylurea</td>
</tr>
<tr>
<td>Acarbose, Miglitol</td>
<td>Alpha-Glucose Inhibitor</td>
</tr>
<tr>
<td>Nateglinide, Repaglinide</td>
<td>Meglitinide</td>
</tr>
<tr>
<td>Pioglitazone, Troglitazone, Rosiglitazone</td>
<td>Thiazolidinedione</td>
</tr>
<tr>
<td>Sitagliptin</td>
<td>Dipeptidyl peptidase 4 inhibitor</td>
</tr>
</tbody>
</table>

Table – 1 Drugs considered in this study.

Clinical Factors Associated with Initial Choice of Therapy
Clinical factors associated with the choice of initial therapy were found by fitting logistic regression models with an L1 regularization, using drug and disease codes prior to the patient’s index time to predict the choice of first line therapy. This analysis was done using the glmnet package in R [14].

Efficacy of Practice-Based Treatment Pathways
The relative efficacy of pairs of treatment pathways were assessed using Cox proportional hazard models adjusted for age and gender. Patients on a specific treatment pathway were considered only if they had a follow up lab value reporting HbA1c profile after the index date. Since the number of HbA1c measurements...
following their index date differed among patients, we considered only the first lab value after the index date. Four levels of HbA1c (6.5%, 7%, 7.5% and 8%) were selected as target thresholds for clinically relevant “control” of HbA1c. For a given target threshold, the time to event was defined as the time at which the first measured HbA1c after the index time was below the threshold. Patients were considered censored if their first reported HbA1c measurement was above the threshold. The hazard ratios and 95% confidence intervals were then computed using cox regression, adjusting for both age and gender of patient.

Results

Practice-Based Treatment-Pathways
The cohort consisted of 6121 patients; most were between 40 to 80 years of age. The gender distribution of patients was observed approximately even, with 2989 females and 3132 male. The ethnicity of the cohort was varied, with Asian (n = 1277), White (n = 2517), Native Hawaiian/other Pacific Islander (n = 168), Black/African American native (n = 357) and American Indian (n = 19) patients, and 1783 patients with unknown ethnicity.

![Figure 1](image)

Figure 1 a) Criteria for cohort construction, b) distribution of patients based on their initial prescription and c) distribution of patients with metformin as initial prescription followed by another drug.

Treatment-pathways were constructed by following the medication trajectory of each patient starting from their respective index dates. Figure 1b illustrates the distribution of first line therapy choices. As expected,
metformin was the predominant first line therapy. However, other drugs such as glipizide, glyburide, pioglitazone and sitagliptin were also used as first line therapies, recapitulating the findings of the OHDSI study. Figure 1c illustrates the distribution of patients who were prescribed metformin as their first prescription but were then prescribed other drugs over a period of time reflecting the choices of second-line treatments. Overall, the trends of initial prescription pattern were consistent with the earlier analysis by OHDSI [12].

Clinical Factors Associated with Choice of First Line Therapy
We searched for clinical factors that are able to explain the choice of first line therapy by fitting L1 regularized logistic regression models [15] to predict first line therapy choices based on the drug and disease codes occurring in the medical record prior to the index time for each patient. Due to the small number of patients in most treatment pathways, this analysis was carried out only for two pairs of initial drug choices: metformin versus glipizide, and metformin versus pioglitazone. Identified factors for each of these treatment pathways are shown in Table 2 below.

<table>
<thead>
<tr>
<th>Initial Choice</th>
<th>Initial Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metformin or Glipizide</td>
<td>Metformin or Pioglitazone</td>
</tr>
<tr>
<td><strong>End Stage Renal Disease</strong></td>
<td><strong>Glucose</strong></td>
</tr>
<tr>
<td><strong>Cefepime</strong></td>
<td><strong>Coronary atherosclerosis of native coronary artery</strong></td>
</tr>
<tr>
<td><strong>Tacrolimus</strong></td>
<td><strong>Diabetes mellitus without mention of complication, type-2 or unspecified type, not stated as uncontrolled</strong></td>
</tr>
<tr>
<td><strong>Diabetes with renal manifestation, type-2 or unspecified type, uncontrolled</strong></td>
<td><strong>Pulmonary collapse</strong></td>
</tr>
<tr>
<td><strong>Esophageal varices in disease classified elsewhere, without mention of bleeding</strong></td>
<td><strong>Acetaminophen</strong></td>
</tr>
</tbody>
</table>

Table 2 Clinical factors predictive of initial choice of therapy. The listed factors are all predictive of the use of glipizide or pioglitazone instead of metformin as first line therapy.

Factors identified in the first comparison, metformin vs glipizide, are consistent with prior knowledge about adverse events associated with metformin. For instance, the choice of glipizide as first line therapy was associated with manifestations of chronic kidney diseases and liver disorders. This is consistent with earlier observations that metformin may induce severe lactic acidosis, which can impair kidney function [16] and thus would be avoided in patients with kidney disease. Another interesting factor is use of the antibiotic cefepime. Earlier investigations suggest a possible drug-drug interaction between cefepime and metformin resulting in reduced efficacy of metformin [17]. On the other hand the factors such as “cirrhosis of liver without mention of alcohol” were predictive of metformin use, consistent with studies showing that metformin is beneficial to diabetes patients with cirrhosis [18].

Similarly, factors associated with initial choice of metformin vs pioglitazone are also broadly consistent with prior biomedical knowledge. For instance, “coronary atherosclerosis of native coronary artery” was predictive of first line use of pioglitazone, consistent with work suggesting that pioglitazone may protect against adverse cardiac events in patients with impaired glucose tolerance [19]. Altogether the identified factors suggest that deviations from the recommended use of metformin as first line treatment for T2DM are largely driven by comorbidities and concerns about adverse events.

Efficacy of Treatment Pathways
The efficacy of pairs of contrasting treatment pathways were assessed by fitting Cox proportional hazard models against the outcome of lowered HbA1c levels (Figure 3). These models compare the rates of outcomes of interest over time in two contrasting groups. We compared five pairs of treatment-pathways. The first three are alternative first line therapy choices (metformin vs glipizide, metformin vs pioglitazone, and metformin vs sitagliptin), while the last two reflect second line treatment choices after metformin (metformin to glipizide vs metformin to pioglitazone, and metformin to pioglitazone vs metformin to sitagliptin). For each of these pairs of treatment-pathways, we computed Cox proportional hazard ratios adjusted for age and gender against a series of outcomes reflecting different levels of HbA1c control – 6.5%, 7.0%, 7.5% and 8.0%. A hazard ratio greater than one means that the treatment pathway highlighted in red text is more effective than the reference treatment pathway (in black text).

As expected, metformin as a first line therapy appeared to be quite efficient in lowering HbA1c compared to glipizide and pioglitazone (Figures 3a and b). Metformin as a first line therapy also showed no significant reduction in the HbA1c profiles in comparison to sitagliptin (Figure 3c), suggesting sitagliptin might be a useful initial therapy controlling the HbA1c profiles to desired levels in situations where metformin is contraindicated due to comorbidities.
The comparisons of treatment pathways with a second line drug suggest that both pioglitazone and sitagliptin are superior to glipizide when added on following metformin, and that sitagliptin is more effective than pioglitazone (Figure 3d and e). If these results hold in other datasets, they suggest that sitagliptin should be considered over glipizide and pioglitazone as second line treatment.

**Discussion**

The choice of second-line treatment for T2DM is not evidence-based due the lack of randomized controlled trials assessing their relative efficacy. Treatment pathways for T2DM derived from observational data in EMRs have been used to elucidate what actually happens in routine clinical practice. We have shown that further analysis of such treatment pathways can also provide insight into why alternative treatment pathways are taken, as well as their relative efficacy. Consistent with work from OHDSI, we found that there was considerable variation from current guidelines in choosing the first line treatment. We then took the next step and identified clinical factors underlying the choice of first line therapy. These factors are consistent with known adverse effects of metformin, as well as comorbidities. Furthermore, analysis of the comparative effectiveness of different treatment pathways suggest that sitagliptin may be as effective as metformin as a first line therapy, and could be an effective choice as second line therapy in cases where monotherapy with metformin fails.

The present analysis is based only on data from Stanford Hospital, which is limited in its population coverage and hence can introduce biases that limit the generalizability of our findings. Furthermore, the strict inclusion criteria regarding pre- and post-index time history result in a relatively small cohort that renders matching on co-morbidities and other patient features before fitting the hazard model problematic. Thus, our results may also exhibit bias due to confounding. It is thus important to replicate this study across multiple sites and in larger cohorts to build generalizable recommendations.

Our initial analysis can be scaled up within the OHDSI research community. Expanding the scale of our work to include data from many different sites may provide further insights into the practice of diabetes management, complement the design of comprehensive clinical trials and help fulfill the goals of learning health-care systems.

**Conclusion**

We have demonstrated the use of observational data from EMRs to address outstanding problems in the management of T2DM that lack evidence from RCTs. These problems include the factors underlying deviations from current guideline recommendations on first line therapies and the relative efficacy of both first and second line therapies. Our approach is easily scalable to incorporate data from any site that has adopted the OHDSI common data model.

**REFERENCES**


8. Association AD. Standards of Medical Care in Diabetes—2016. Diabetes Care 2016;39(Supplement 1)


Melinda – A custom search engine that provides access to locally-developed content using the HL7 Infobutton standard

Yik-Ki J. Wan, BS1, Catherine J. Staes, BSN, MPH, PhD1,

1Department of Biomedical Informatics, University of Utah,
Salt Lake City, Utah, USA

Abstract

Healthcare organizations use care pathways to standardize care, but once developed, adoption rates often remain low. One challenge for usage concerns clinicians' difficulty in accessing guidance when it is most needed. Although the HL7 ‘Infobutton Standard’ allows clinicians easier access to external references, access to locally-developed resources often requires clinicians to deviate from their normal electronic health record (EHR) workflow to use another application. To address this gap between internal and external resources, we reviewed the literature and existing practices at the University of Utah Health Care. We identify the requirements to meet the needs of a healthcare enterprise and clinicians, describe the design and development of a prototype to aggregate both internal and external resources from within or outside the EHR, and evaluated strengths and limitations of the prototype. The system is functional but not implemented in a live EHR environment. We suggest next steps and enhancements.

Background and Significance

Among industrial nations, the US ranks highest in health care costs and last in quality, efficiency and patient outcomes. In a recent White House report, the US President attributed these inefficiencies to the lack of coordination of care, over-treatment, and the failure to adhere to best practices. For example, one study shows that 53% of patients who have undergone cataract surgery still received unnecessary preoperative testing even though guidelines recommending against such testing were published 20 years ago. To address these problems, healthcare organizations began developing care pathways to standardize best practices. However, difficulties with accessing these care pathways at the point of care impede adoption. Therefore, organizations use a variety of knowledge management solutions for delivering resources and guidelines, ranging from storing static documents in a shared network location, customizing off-the-shelf products (such as Microsoft SharePoint), configuring internal wiki pages, and developing custom mobile applications. While these approaches are better than the archaic paper-in binder approach, none allow users to access documents from within electronic medical records where care pathways and support for clinical decision-making are needed most.

The difficulty of retrieving relevant clinical guidance at the point of care is a well-documented problem. To address this problem, the HL7 Context-Aware Information Retrieval (e.g., Infobutton) Standard was adopted as a Meaningful Use Requirement. The standard specifies the usage of controlled terminologies, and query and reply formats. Using a service-oriented architecture, the basic workflow is as follows: 1) a standardized query is submitted as a knowledge request to the Infobutton Manager; 2) the Infobutton Manager interprets the clinical context within the request and matches against available resource profiles; 3) once matched, the Infobutton Manager replies with the Uniform Resource Locator (URL) to any matched resources (see top section of Figure 1). Depending on the receiving application, the user may be presented with the resources' URL or the content fully rendered. Although Infobutton is designed to be flexible for local adoption, most accessible resources are developed externally and published by third-party content providers. Local content is not accessible through Infobutton until a resource context profile is defined by indexing with standard terminology. OpenInfobutton, an open-sourced implementation of the Infobutton standard, includes a Librarian Infobutton Tailoring Environment (LITE) to ease the creation of context profiles.

Information retrieval problem at UUHC

University of Utah Health Care (UUHC) is an academic, regional healthcare system. UUHC consists of both inpatient hospitals and outpatient clinics serving patients in Utah and surrounding areas. Since 2014, Epic has been the electronic health record (EHR) in the inpatient and outpatient settings. The Epic system includes an Infobutton solution called ClinKB. OpenInfobutton serves as the Infobutton Manager for Epic's Infobutton requests. Users can initiate an Infobutton query by clicking an icon next to a discrete patient data field such as a diagnosis, lab result, or...
medication. The query is then submitted to the OpenInfobutton. OpenInfobutton then aggregates and returns resources from third-party content providers such as UpToDate, Micromedex and Karma. Like most Infobutton implementations, accessing an Infobutton requires that a provider first be signed into Epic, search for a patient's record, find the relevant discrete data with an Infobutton icon, and then click on the icon to query for the references. To date, content developed locally is not part of the result set. For locally developed content, UUHC uses Microsoft's SharePoint, which also serves as a collaboration space for all workforce teams (see Figure 1). Naturally, searches in SharePoint only return locally developed content.

![Figure 1. Current UUHC Information access paths for internal and external resources](image)

Many approaches have been tried to meet the information needs at the point of care. Early on, static documents were compiled into physical binders and distributed to each nursing unit within the medical facilities. Access and updates proved to be cumbersome. A less physically-bound approach was implemented by adding static links to the EMR's toolbar menu. However, the number of searchable links became unmanageable as the number of documents grew, and the linked guidance became outdated. Another approach used links within Epic to redirect users to SharePoint. While this method did allow providers to search for local content, the search results often included clinically less relevant documents. Another approach was adding resource links to Best Practice Alerts (BPA) within Epic. This approach meant resources were only accessible after providers have performed specific steps that triggered the BPA. Often decisions were already made and alerts were ignored.

Using Infobutton to access care pathways poses another challenge. Since Infobutton is designed to serve relevant resources within clinical contexts, many resources are only available after the provider has established a clinical context by making some clinical decisions. Conversely, care pathway documents often are needed before decisions are made. One organization added a search box to their Infobutton result page for their providers to find resources beyond the scope of the established context. They have found that providers used this search box more than half the time. Their finding demonstrated that the need for a more flexible search was greater than expected. Similarly, at UUHC, even with the availability of Infobutton and SharePoint, providers still have difficulty accessing institutional guidelines and content. A new approach is needed to leverage the relevance of Infobutton search results while providing the flexibility of free searches for both internal and external content.

**Objective**

Our long-term goal is to utilize the Infobutton standard and provide users access to both local and external content at the point of care even before the establishment of a patient's context. Our specific objectives for the work described in this paper were to a) define major requirements to provide users more flexible access, b) design and develop a prototype solution, and c) evaluate strengths and limitations of the prototype and propose next steps for evaluation and development.

**Define major requirements**
During 2015, a team of informatics graduate students collaborated with a multidisciplinary team of clinicians and IT hospital operations staff responsible for the EHR and enterprise-wide document control. The team focused on the challenge of implementing care pathways and communicating recommended guidance. After performing user testing of clinical alerts associated with the care pathway guidance, the team reviewed the literature and started to document requirements particularly focused on making information available at the point of care. The team reviewed the current knowledge workflow within the organization and determined that a custom solution may meet that need. After several meetings with stakeholders responsible for document control, quality control, and clinical care, the following major requirements were identified:

**Functional requirements**

- **Providers should be able to access care pathways within EHR:** Since providers began using Infobutton to access references, the solution should allow providers access to locally-developed care pathway documents using Infobutton. Infobutton access makes it easier for providers to search for these documents.

- **Providers should be able to generate custom Infobutton queries at the point of care:** The effectiveness of Infobutton depends on the precision of the resource context profiles. However, because the context is defined with high specificity, the reference often can only be accessed after a matching context is established. Of course, generalizing the context profile would return more results, but it would also decrease precision. Instead of generalizing the context profiles, the solution should enable the user to customize the query.

- **Providers should be able to search resources by free-text:** One of the difficulties in formulating Infobutton queries is using relevant controlled terminologies. Free-text search shifts the burden of term selection from the user to the system. The solution should allow providers to perform free-text searches.

- **Providers should be able to access both internal and external resources from a single tool:** Outside of Infobutton, UUHC providers use SharePoint to search for internal documents. For external references, providers must remember the appropriate content provider's website. Since this multi-portal approach has been a hurdle in care pathway implementation, the solution should search and aggregate both internal and external content, forming a single point of access.

- **Providers should be able to access the tool from anywhere:** Allowing access to reference materials independent of an EHR is important. In fact, most medical reference content brokers include searches on their websites or mobile applications. Conversely, searching for resources indexed within Infobutton requires multiple logins to access the organization's EHR through a secure network. Separating protected health information from openly available references allows searches from any Internet-enabled device and may lead to wider dissemination of best practices. Therefore, the solution should be accessible from inside and outside of the EHR.

**User Interface design requirements**

An effective search engine should be simple to use, provide useful feedback and display accurate results with meaningful metadata to help users select the most relevant results. With these goals in mind, we identified the following user interface (UI) related requirements:

- **The UI should be fast, simple and easy to use:** Ideally, users should be able to use it without explicit instruction. The layout of the website should direct users to the entry point of the workflow. The user interface should be responsive, intuitive and fast in returning results.

- **The solution should provide meaningful feedback:** The current appearance of the Infobutton icon does not indicate if a locally developed resource exists. Anecdotally, some users hesitate to use Infobutton because it may detract from their workflow without gaining any useful information. An effective search engine should provide feedback even before the query is issued. Since providers and the institution may consider internal content to be more relevant, the solution should give an indication if the query will return any local content.

- **The system should provide customizable autocomplete ranking for more relevant results:** Autocomplete had been the most used Google Search feature. The selection of the search term that affects the accuracy of search results is heavily influenced by the entry's ranking. With the promotion of a certain autocomplete suggestion, UUHC would be able to deliver relevant resources more accurately.
Prototype development

To build within the Infobutton ecosystem, two parallel efforts were required: 1) identify resources requested by clinicians in a primary care setting that are stored in a location accessible by URL, and then index the resources' URLs using LITE and creating their resource profiles; and 2) develop a custom search engine to initiate an Infobutton knowledge request and present the resulting URLs of the resources.

Identify and index resources using Librarian Infobutton Tailoring Environment (LITE)

Resources were selected two ways. We included four care pathway diagrams recently developed by UUHC clinicians for knee pain, concussion, and seizures. We also included 14 resources selected by a lead pharmacist and a nurse manager in a UUHC community clinic. These resources were accessible by URL and included because they were either frequently used or desired to be accessible. Most of the sample resources identified (n=16) currently reside in UUHC's internal SharePoint server. Two are external resources (Table 1).

Table 1. Description of resources curated and indexed using the UUHC Infobutton LITE application

<table>
<thead>
<tr>
<th>Title of Resource</th>
<th>Resource Type</th>
<th>Resource Format</th>
<th>Internal or External</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Knee Pain Pathway - traumatic</td>
<td>Care Pathway</td>
<td>PDF</td>
<td>Internal</td>
</tr>
<tr>
<td>Acute Knee Pain Pathway - non traumatic</td>
<td>Care Pathway</td>
<td>PDF</td>
<td>Internal</td>
</tr>
<tr>
<td>Concussion Pathway Diagram</td>
<td>Care Pathway</td>
<td>PDF</td>
<td>Internal</td>
</tr>
<tr>
<td>Seizure Care Pathway Algorithms</td>
<td>Care Pathway</td>
<td>PDF</td>
<td>Internal</td>
</tr>
<tr>
<td>Chest Pain Workflow - Community Clinic</td>
<td>Care Pathway</td>
<td>DOC</td>
<td>Internal</td>
</tr>
<tr>
<td>Guideline: Kidney Transplant Inpatient Management</td>
<td>Care Pathway</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Headaches in Pregnancy</td>
<td>Care Pathway</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Guideline: Pulmonary Embolism Care Pathway Algorithms</td>
<td>Care Pathway</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Guideline: Inpatient Thrombosis Service Pharmacist</td>
<td>Care Pathway</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Hypertension: Clinician Guideline Summary</td>
<td>Care Pathway</td>
<td>HTML</td>
<td>External</td>
</tr>
<tr>
<td>Adult Blood Pressure Measurement</td>
<td>Clinical Task</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Intravenous (IV) Medication Policy and List</td>
<td>Clinical Task</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Warfarin Dosing</td>
<td>Medication</td>
<td>HTML</td>
<td>External</td>
</tr>
<tr>
<td>Alternatives to nadolol</td>
<td>Medication</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>CDC Hand Hygiene Patient Brochure</td>
<td>Pt Education</td>
<td>PDF</td>
<td>Internal</td>
</tr>
<tr>
<td>MyChart FAQ</td>
<td>Pt Education</td>
<td>DOC</td>
<td>Internal</td>
</tr>
<tr>
<td>Medication Refill</td>
<td>Pt Education</td>
<td>HTML</td>
<td>Internal</td>
</tr>
<tr>
<td>Patient Complaint Process</td>
<td>Pt Education</td>
<td>PDF</td>
<td>Internal</td>
</tr>
</tbody>
</table>

The resources represented an array of references requested by clinicians working in outpatient primary care settings. Ten of the resources were care pathways and clinical guidelines that could be used to guide clinicians in treatment planning. Four resources were patient education materials that could be used in various contexts. Two documents addressed specific clinical tasks (i.e., blood pressure measurement and IV line care) that may be useful as a reference for interns and students outside of a single patient’s context. Two additional documents related to
medication administration and guidance that may be useful before a physician places an order. Currently, the 16
resources that reside in SharePoint can only be accessed by URLs after user authentication has been established, and
after searching the site for the resource.

Independently, two informatics graduate students assigned standardized codes that should be used to index the
resources in LITE according to the HL7 specification, including for example SNOMED CT codes for the main
search criteria and other code systems for the performer, encounter, and task context properties. The students had
access to the specification and experience with clinical-related controlled terminologies through their informatics
course work. The codes were then reviewed by an informatics research faculty member (author CS) with extensive
experience in controlled terminologies, followed by discussions with the two students to reconcile differences. After
the list of codes was finalized, the two students created context profiles for each resource using LITE. Note that
LITE has functionality that allows the user to select the major concept heading for the "MainSearchCriteria"
element, and will then populate additional MainSearchCriteria with the children of the major concept heading based
on hierarchies in SNOMED CT. For those resources that were not related to a particular disease process (such as the
Patient Complaint Process Form), the "MainSearchCriteria" element was left unpopulated.

Design and Develop the Prototype

The design of the custom search engine was based upon service oriented architecture and consisted of three parts:

1. A front end application allowed for user input and submitted the input to the server for suggestions.
2. A backend server received user input, queried the database, and replied with appropriate suggestions.
3. A custom database contained all the available resource titles and controlled terminologies for suggestions.

The prototype was named Melinda. Its focus was to help a user easily formulate an Infobutton request using an
autocomplete function and submit the request to the Infobutton Manager. Figure 2 illustrates the system's
components and their interactions with the Infobutton Manager. The front-end user interface was developed using
Angular. Django was chosen as a web framework for the backend server because of its ease of in prototyping. A
SQLite database contained the resource titles and the set of standardized terms and associated codes used for
matching by Melinda. Melinda is hosted by Heroku as www.Melinda.tech

![Figure 2. Proposed Design of Infobutton Search Engine Information Process](image)

The terms used for user-input matching came from two sources. First, we downloaded the "CORE Problem List
Subset of SNOMED CT" from the National Library of Medicine UMLS Terminology Services and loaded the
6,166 concepts into the matching table. We selected this source because the complete list of terms in SNOMED CT
is too large and many terms would not be relevant. This subset of most used terms to represent clinical problems
seemed sufficient to seed the system for the prototype. Second, we used a new service recently-developed for our
use in the OpenInfobutton Manager that allowed Melinda to directly access and download the standard codes
assigned for the MainSearchCriteria in the resource profiles. We loaded these codes into the matching table as well.
Therefore, each row in the matching table contained the term, the code, and the Object Identifier (OID) for the corresponding code system.

The matching process has three major steps. Step 1: As a user starts typing in the text box, the front-end application takes the user’s input and submits the input for matching. Step 2: The server matches the input against the terms in the matching table as well as the titles of the indexed resources. The matching is done by tokenizing the user’s input into list of words and matching them with the terms in the matching table. Step 3: The server replies with suggestions consisting of terms and/or resource titles. The replies with terms include the code and code system required for an Infobutton request. The replies with resource titles include the resource URLs for direct Ajax calls. The suggestion ranking is customizable in Melinda’s database, allowing an enterprise to prioritize internal versus external resources. Any entry that is associated with locally-developed content is displayed with an icon as shown in Figure 3.

![Figure 3](image)

**Figure 3.** Melinda’s user interface as user first types in the text box

The response process has several steps depending on the user’s selection. If the user selects a term from a standard terminology, Melinda will build an Infobutton request with the associated code and submit the request to the Infobutton Manager. Then, Melinda parses the resulting JSON object from the reply, sorts the URLs by the source of the content, and displays the results. In contrast, if the user selects a resource title, Melinda extracts the URL and displays the link directly to the user, bypassing the Infobutton request-reply workflow altogether. The results are displayed to the user organized by the source of the resource. The order of the sources in the display is a customizable feature. For each result entry, the title, URL and last update date are displayed as shown in Figure 4.

![Figure 4](image)

**Figure 4.** Melinda’s display of results, showing the metadata including title, URL, and last update date
Strengths and Limitations of the current prototype

Strengths

- Users have a single point of access for both internal and external content: Once internal content was indexed within the Infobutton Manager, relevant (i.e., matching) internal resources were included in the Infobutton reply along with relevant external resources in a single page. Users were able to access both internal and external resources from a single application.

- Users can perform search outside the current patient context: With Melinda, clinicians now can perform any searches without the need to first establishing a patient context.

- Users can perform custom searches at the point of care: Melinda's autocomplete suggestions allows providers the flexibility to search any resource, even those that were without MainSearchCriteria populated in their profiles. This added flexibility allows content authors to define the resource profiles as precisely (or as generally) as needed without hindering access to a resource.

- Users can access Melinda anywhere: Since Infobutton context profiles were accessible publicly, and Melinda was hosted on a public domain, the query result was accessible to anyone without the burden of logins. Moreover, Melinda can detect device display size and provide navigation layout appropriate for small mobile devices, tablets, or full desktop display. Screenshots of the interface rendered on a mobile device are shown in Figure 5.

- Finer control of information distribution for the organization: Melinda's autocomplete suggestion and result rankings are both customizable. The sorting order allows UUHC to sort most relevant and targeted information to the top, thus allowing providers easier access to common organization-specific resources. Furthermore, many accreditation programs, such as ISO9001, require healthcare organizations to establish a policy update and access procedures. This approach may be part of such an established system.

- Better user experience: The three principles of effective search engine design guided Melinda's UI development. For simplicity, the single search box design was chosen to minimize confusion and distraction. Second, since locally-developed guidelines were more likely to be considered than external resources, the suggestion list's special icon allows users to see which searches would return locally-developed content before the search was performed. Finally, the title, URL and updated date and time associated with the resources help the user determine the relevancy of the result.

Limitations

- Possible user confusion if similar terms from different coding system return different resources: To ensure adequate matches, system administrators may include terminologies with overlapping domains of clinical content. As a consequence, users may be presented with a set of similar terms that return different resources when used in a query. For example, if terms from SNOMED CT and ICD-10 were loaded into the matching
table and a user inputs "knee pain", both "Knee Pain (findings)" from SNOMED CT and "Pain in unspecified knee" from ICD10 would be presented. Depending upon the MainSearchCriteria indexed in the resource profiles, each term may return different results.

- **Lack of full integration into EHR:** Currently, the prototype was only a proof of concept. While it was functional, the lack of full integration with Epic means access to the tool remains isolated from Epic. Within Epic, access to care pathways continues to be Infobutton icons or the dedicated link to SharePoint. The theorized benefits of custom search by Melinda within the EHR can only be realized after UUHC Epic's ClinckKB environment is configured to direct requests to Melinda.

- **Limited context tailoring:** Separating the Infobutton queries from protected health information introduced one major drawback: the user now must re-establish the context manually for Infobutton. When forming the request, even though Melinda could search by the most used Infobutton resources criteria (i.e. mainSearchCriteria), the absence of other contextual information may diminish the relevance of the search results. Soliciting for these additional parameters would require additional user input which the current simple UI design does not yet accommodate.

- **Lack of free-text search:** The quality of Infobutton queries rests on the use of standard terminologies, particularly a standard codes' Universally Unique Identifiers (UUIDs). To this end, Melinda supplied standard terms (with associated codes) for the user to select. In turn, Melinda converted the term to its corresponding UUID to make a request. In essence, the search box is not a true free-text search box but an expanded picklist. This approach limits the flexibility of free text searches, a feature commonly used by other external content providers. Future enhancements could address this need.

Even with these limitations, adding Melinda to the Infobutton framework may expand the power of Infobutton to other clinical decision support solutions. Melinda met many of the initially-identified requirements (Table 2).

**Table 2. Requirements of the Infobutton search engine and their fulfillment status by the prototype.**

<table>
<thead>
<tr>
<th>Requirements:</th>
<th>Requirement Source: (Literature, UUHC problem analysis or Both)</th>
<th>Is the requirement met in the current prototype?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functional requirements - The User should be able to:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...access care pathways within EHR</td>
<td>UUHC Analysis</td>
<td>Not yet implemented in production Infobutton manager</td>
</tr>
<tr>
<td>...generate custom Infobutton queries at the point of care</td>
<td>Both</td>
<td>Partial - only MainSearchCriteria is used</td>
</tr>
<tr>
<td>...search resources by free-text</td>
<td>Both</td>
<td>Not Met</td>
</tr>
<tr>
<td>...access both internal and external resources from a single tool</td>
<td>UUHC Analysis</td>
<td>Met</td>
</tr>
<tr>
<td>...access the search tool from anywhere</td>
<td>Both</td>
<td>Partial - accessible outside of EMR only.</td>
</tr>
<tr>
<td><strong>User Interface requirements - The system should:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...be fast, simple and easy to use</td>
<td>Both</td>
<td>Not yet evaluated</td>
</tr>
<tr>
<td>...provide meaningful feedback</td>
<td>Literature</td>
<td>Met</td>
</tr>
<tr>
<td>...provide customizable autocomplete ranking</td>
<td>Literature</td>
<td>Met</td>
</tr>
</tbody>
</table>

**Future work**

**Evaluation**

The usability and functionality of Melinda needs to be evaluated more formally. First, Neilson' usability heuristic criteria should be applied to evaluate Melinda's usability before provider end user feedback. After most of the usability defects and issues are addressed, we should then perform user studies with five to 10 providers who
represent the target audience for this application. For example, the user study should include physicians, pharmacists, and nurses in the outpatient primary care setting during the first phase of testing. These evaluations will provide input for enhancements and guide further evaluation plans before Melinda can be implemented in a production setting. Once Melinda can be implemented in a production environment, on-going monitoring should be carried out to assess usage, satisfaction, system response times, and the relevance (i.e., predictive value positive) of terms presented to the user based on selections made by the user. Ongoing feedback may be solicited with the use of a simple feedback button.

**Enhancements**

The following future enhancements can be built upon the current framework:

- **More context awareness:** The Infobutton Standard allows for use of other context information that Melinda is currently not gathering. With additional context information, the system may be able to learn to map frequent user input to the appropriate standard terminology (e.g., ‘K’ for Potassium level). Of course, such features will require a machine learning middleware and an efficient user interface with a feedback loop.

- **Encapsulate Melinda into a web service:** Minor customization of Melinda could convert the web page to a web service that passes Infobutton requests along to Infobutton Manager. As a web service, Melinda could also replace the current Infobutton responder page and unify the interface between Infobutton icon access and search page access. As a web service, Melinda could integrate more fully and easily with an EHR.

- **Security:** While the resources at UUHC were protected behind SharePoint's user authentication, the indexes and metadata could be accessible without user authentication, allowing open access to Melinda. If security were needed, single sign-on technology, such as LDAP or OAuth might be beneficial to free the user from having to remember yet another secure identity.

- **Individual accounts:** The framework design could also be leveraged as a mobile app. If user accounts were implemented, a mobile app might be built to include advanced features such as push notification and user feedback. Melinda could query a provider's schedule and calendar to present needed resources as a custom reading list based on chief complaints of patients upcoming in their clinic. User's interaction tracking for dynamic page ranking may be added as a well.

**Conclusion**

We identified key requirements for a system to support the clinician’s needs to access clinical guidelines that were developed internally and externally both from within and outside an individual patient’s clinical record (i.e., both with and with clinical context). A prototype was developed to demonstrate the feasibility and to study further the system and user requirements for full implementation. While the benefits may seem promising, the limitations we uncovered would need to be addressed in future development. This research addresses a major gap in the current strategy for accessing relevant and updated information important for clinical care.

**Acknowledgment**

We thank Guilherme Del Fiol for providing insight and expertise about the use of Infobutton and access to LITE, Bruce Bray for providing insight into the clinical setting workflow, Andrew Iskander for developing a web service that allowed Melinda to query Infobutton Context Profiles, Bret Heale and Stacey Slager for creating the resource profiles, and UUHC Value Driven Care Team for selecting the sample reference resources.

**References**


13. Federal Register Volume 77, Number 45 (Wednesday, March 7, 2012), Pages 13832–13885


Term Coverage of Dietary Supplements Ingredients in Product Labels

Yefeng Wang¹, Terrence J. Adam, RPh, MD, PhD¹,², Rui Zhang, PhD¹,³

¹Institute for Health Informatics; ²College of Pharmacy; and ³Department of Surgery; University of Minnesota, Minneapolis, MN

Abstract

As the clinical application and consumption of dietary supplements has grown, their side effects and possible interactions with prescribed medications has become a serious issue. Information extraction of dietary supplement related information is a critical need to support dietary supplement research. However, there currently is not an existing terminology for dietary supplements, placing a barrier for informatics research in this field. The terms related to dietary supplement ingredients should be collected and normalized before a terminology can be established to facilitate convenient search on safety information and control possible adverse effects of dietary supplements. In this study, the Dietary Supplement Label Database (DSLD) was chosen as the data source from which the ingredient information was extracted and normalized. The distribution based on the product type and the ingredient type of the dietary supplements were analyzed. The ingredient terms were then mapped to the existing terminologies, including UMLS, RxNorm and NDF-RT by using MetaMap and RxMix. The large gap between existing terminologies and ingredients were found: only 14.67%, 19.65%, and 12.88% of ingredient terms were covered by UMLS, RxNorm and NDF-RT, respectively.

Introduction

According to the National Health and Nutrition Examination Survey (NHANES) 2003-2006, 53% of U.S. adults took at least one dietary supplement, most of which were multivitamin and multi-mineral supplements. Recently, the National Health Statistics Reports indicated that about 40% of Americans use some form of complementary and alternative medicine (CAM) and that non-vitamin/non-mineral dietary supplements were still the most commonly used complementary and alternative health approach. The prevalence of supplement use was estimated to be 69% in 2011 when occasional and seasonal use was taken into account.

Dietary supplements are typically used to complement conventional medicine with the goal of achieving better healthcare outcomes; however, dietary supplements were found to result in 23,000 emergency room visits yearly in the U.S. according to the study conducted by Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC). Moreover, one out of four people taking a prescription medicine also uses an herbal supplement, increasing the possibility of drug-supplement interactions (DSIs). For example, warfarin can interact with supplements such as Panax ginseng and Gingko biloba, which can lead to severe adverse effects such as spontaneous postoperative bleeding. In particular, supplements are increasingly used by patients diagnosed with cancer to help strengthen their immune system and ease the side effects of treatments. Unfortunately, our ability to identify adverse effects from herb and dietary supplements and their reactions with conventional Western medications are currently limited and the reports on such interactions occur infrequently in clinical practice. Therefore, it is necessary to gather information on the ingredients in these supplements to facilitate medication safety efforts.

New drugs are typically tested for their efficacy and toxicity before market approval. However, U.S. Food and Drug Administration (FDA) regulates dietary supplements differently from conventional food and drugs under a separate regulation called Dietary Supplement Health and Education Act of 1994 (DSHEA). DSHEA requires appropriate labeling of dietary supplements. These labels contain rich information such as suggested use, ingredients, product indication, target population, and other necessary safety precautions. In particular, the dietary supplement labeling guideline requires all dietary ingredients to be listed on the label, and the synonyms for the dietary ingredients can be used. Such information provides a great resource to analyze the ingredients of dietary supplements from the perspective of terminology.

To support effective mining of product labels information for DSIs and supplements adverse effects studies, it is vital to understand the representation of supplements and ingredients in dietary supplement labels as well as the coverage of standard biomedical terminologies for supplements and their ingredients. One possible reason for the lack of
extensive studies on DSIs is an absence of a standard and accepted terminology for dietary supplements. In a recent study comparing terms with different resources, we found that none of five major online databases covered all supplement terms\(^8\). Another prior study also evaluated the supplement term coverage in both medication lists and clinical notes in the electronic health record (EHR)\(^9\). To the best of our knowledge, the investigation of supplement term representation and coverage on dietary supplement labels is limited and deserves further investigation. In this study, we sought to evaluate terminology coverage of data elements in supplement product labels by cross-validating online supplement databases with EHR patient data. We also sought to investigate the adequacy of standard terminologies for representing supplements, especially their product ingredients, based on existing product labels.

**Methods**

We extracted supplement product information and their ingredients from the Dietary Supplement Label Databases (DSLD), and then normalized and formed a comprehensive list of ingredients followed by mapping them to standard terminology, including UMLS, RxNorm, and NDF-RT. The term coverage in these terminologies was investigated. In our assessment, we compared the term overlap among three resources, including supplement labels, online resources, and the EHR medication list. We first introduce DSLD, MetaMap, and RxMix which are used in this study, followed by the details of methods.

**DSLD**

The DSLD is created by the National Institutes of Health (NIH) Office of Dietary Supplement (ODS) and U.S. National Library of Medicine (NLM). The database collects full label contents from dietary supplement products including both currently available products as well as products which are no longer on the market in the U.S., as well as those consumed by National Health and Nutrition Examination Survey (NHANES) participants. Each product in DSLD provides four types of label information: product information (including serving information, product type, supplement form, target groups), dietary supplement facts (e.g., usage, ingredients), label statement (e.g., FDA statement, precautions, and suggested use), and contact information.

**Unified Medical Language system (UMLS) and MetaMap**

UMLS is a repository that integrates over 100 medical vocabularies and provides a unified platform which can be used to develop or enhance applications. Metathesaurus, as one of the three main components in the UMLS, contains over 2 million terms and codes from many different vocabularies such as Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), Current Procedural Terminology (CPT), Logical Observation Identifiers Names and Codes (LOINC), etc. Each concept in the Metathesaurus has a Concept Unique Identifier (CUI), which can be used to map to various dictionaries and compare data from different sources. For this paper, UMLS was used to normalize data from different sources (i.e., supplement ingredients).

To map the biomedical texts to the UMLS Metathesaurus, the NLM developed and maintains a natural language processing tool, called MetaMap, which lexically and syntactically analyzes texts and provides a list of mapping concept candidates. We used MetaMap to find the matched UMLS concepts in Metathesaurus. The versions of MetaMap and UMLS Metathesaurus used in this study are 2016 and 2015AB, respectively.

**RxNorm, NDF-RT, and RxMix**

RxNorm provides a standardized nomenclature of clinical drugs by integrating 12 drug vocabularies\(^10\). RxNorm contains names not only for prescription drugs but many over-the-counter drugs on the U.S. market. RxNorm provides the ability to normalize the drug names along with drug dosage, generic name, chemical components, and dosage forms to support the communications between different systems. The version of RxNorm used in this article is “04-Jan-2016”.

National Drug File - Reference Terminology (NDF-RT) is developed by the U.S. Department of Veterans Affairs, Veteran Health Administration. It is used to represent drug characteristics such as ingredients, chemical structure, and additional information about their molecular interactions and kinetics\(^11\). The version of NDF-RT used in this article is “2016.01.04”.

RxMix is a web application integrating various functions from the RxNorm and NDF-RT application program interfaces (APIs) to allow users to create their own workflow to conduct a certain task, such as finding drug class from a drug name\(^12\). In this paper, RxMix was used to map terms to RxNorm and NDF-RT concepts.
Extracting and analyzing product information from DSLD

Step 1: Downloading the data from DSLD. For each supplement in DSLD, the data containing the abovementioned label information are presented in four separate comma-separated values (CSV) files that are available for download. DSLD consists of 45,455 up-to-date records, downloaded through batch processing by Python according to the unique DSLD identifier of each supplement.

Step 2: Extracting the product information of the supplements. The information pertaining to the product name and the product type can be found in the Product Information. DSLD used the LanguaL™ system for the classification of the supplements since LanguaL™ is an “automated method for describing, capturing and retrieving data about food” 13. In total, 12 categories (e.g., single vitamin and single mineral supplement) were applied to label more than 45,000 dietary supplements. It is noteworthy that no generic names for these supplements can be found in the database, and the manufacturers have come up with their unique nomenclature for these mixtures. Therefore, the most useful information in terms of terminology in the product information label is the LanguaL™ supplement type. LanguaL™ assigns the letter “A” to all dietary supplements and a four-digit code following the letter for further classification. The supplement types for each supplement were extracted and the distribution of the supplements in each type was then analyzed.

Normalizing and analyzing supplement ingredient list

Since no generic names could be found for the supplements in DSLD, more emphasis were put on scrutinizing the ingredients of each supplement. Here, we describe how to extract, normalize and analyze ingredients for each supplement.

Step 1: Extracting ingredient names and ingredient types. The information pertaining to the ingredient names and ingredient types was further extracted from Dietary Supplement Facts. Therefore, it is straightforward to iterate through the items in each dietary supplement facts file and extract the relevant information. The number of ingredients in each supplement was also recorded. Then the information were categorized into 20 different types (e.g., vitamin and amino acids) which appeared in the data. Among these, two categories of ingredients (i.e., default and header) were excluded in this study as these types of ingredients are not specifically classified in the database. There are 60,005 items that are belong to this class that was named “default”. The “header” type of ingredients is used to organize the presentation structure of dietary supplement facts on the label, while the information itself does not contain any useful information about the ingredients.

Step 2: Normalizing the ingredient information. After extracting the ingredient raw data, it was discovered that further data cleaning is required before any terminology mapping could be done. We removed information including comment in parentheses, brackets and braces, HTML marker residue from webpage sources left in the CSV files during the transfer process, non-alphanumeric characters and various duplicates for a single ingredient type with regular expression filters. It was found that the different formulations of the same ingredients contributed to the majority of redundant information. For example, the ingredient acai has multiple duplicates in the record: acai extract, acai powder, acai freeze-dried powder, acai concentrate, acai juice, etc. The word “extract”, “powder”, “concentrate” does not contribute to efficient mapping. To eliminate all these duplicates, a keyword list with formulations (e.g., extract, powder, concentrate) was manually summarized and used to preprocess these ingredient list using additional mapping steps. The list of original and normalized ingredients and the Python source code are available at: https://github.com/Schneitzer/Term-Coverage-of-Dietary-Supplements-in-Product-Labels.

Step 3: Analyzing the ingredient information. After normalizing the ingredient information, a list of unique terms and their ingredient types were generated. The normalizing efficiency can also be deduced by comparing the number of items in each set before normalizing and after normalizing.

Mapping ingredient terms to existing terminologies

After normalizing the ingredient data, seven subsets were chosen without loss of generality, i.e., amino acids, animal part or source, botanical, chemical, enzyme, hormone, and vitamin, for mapping.

Step 1: Mapping the list of normalized ingredients to UMLS using MetaMap. The mapping result contains a matching score according to the accuracy of mapping could be measured. A score of 1000 represents a perfect match, while lower scores indicates a partial match. It is possible that a certain ingredient could be mapped to multiple keywords, which will obfuscate its clinical significance due to a partial match. Therefore, only mappings with 1000 matching score were further analyzed in this study.
Step 2: Filtering the list of ingredient information according to the semantic types. The UMLS semantic types of the exact matched concepts were examined as some of those concepts may not refer to dietary supplements. For example, the concept “beta carotene” has two exact match in UMLS Metathesaurus, but only one of them refers to “organic chemical, pharmacologic substance, and vitamin”; the other maps to “Beta carotene Measurement”, which is a laboratory procedure. In addition, although the ingredient information has been preprocessed and normalized, there are still items that may contain long phrases that may produce false positives. Therefore, only a certain subset of semantic types (e.g., “Enzyme, organic chemical”) that are closely related to dietary supplements were manually chosen to filter out the mapping noise.

Step 3: Mapping the list of normalized ingredients from each ingredient type to RxNorm and NDF-RT using RxMix\textsuperscript{12}. The normalized lists were directly uploaded to RxMix workflow for mapping. Two function, \texttt{RxNorm:findRxcuiByString} (with normalized string match parameter selected) and \texttt{NDF-RT:findConceptsByName}, were applied to find the unique concept identifiers in RxNorm and NDF-RT terminologies.

Evaluating ingredient term coverage of terminologies

After the ingredient information was mapped to UMLS, RxNorm and NDF-RT terminologies, the unique concepts were collected although many ingredient terms can be mapped to the same concept although we normalized the ingredients terms described above. The number and percentage of the unique mapped concepts in each terminology were calculated to indicate the unique terms of the ingredient information of DSLD that can be mapped to all the databases included in this research.

Results

Extraction of supplement list from DSLD

All 45,455 supplements from DSLD databases were extracted and classified by 12 LanguaL\textsuperscript{TM} product types. Number and percentage of each product type were shown in Table 1.

Table 1. The supplement representation in DSLD; number, percentages and LanguaL\textsuperscript{TM} product type.

<table>
<thead>
<tr>
<th>LanguaL\textsuperscript{TM} Product Type</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIETARY SUPPLEMENT – COMBINATION/OTHER</td>
<td>16531</td>
<td>36.4%</td>
</tr>
<tr>
<td>DIETARY SUPPLEMENT – HERBAL OR BOTANICAL</td>
<td>8681</td>
<td>19.1%</td>
</tr>
<tr>
<td>DIETARY SUPPLEMENT – NON-NUTRIENT/NON-BOTANICAL SUPPLEMENT</td>
<td>5448</td>
<td>12.0%</td>
</tr>
<tr>
<td>BOTANICAL SUPPLEMENT WITH VITAMIN/MINERAL</td>
<td>5099</td>
<td>11.2%</td>
</tr>
<tr>
<td>DIETARY SUPPLEMENT - VITAMIN</td>
<td>2922</td>
<td>6.4%</td>
</tr>
<tr>
<td>DIETARY SUPPLEMENT - MINERAL</td>
<td>1727</td>
<td>3.8%</td>
</tr>
<tr>
<td>DIETARY SUPPLEMENT – AMINO ACID OR PROTEIN</td>
<td>1558</td>
<td>3.4%</td>
</tr>
<tr>
<td>FATTY ACID OR FAT/OIL SUPPLEMENT</td>
<td>1444</td>
<td>3.2%</td>
</tr>
<tr>
<td>MULTI-VITAMIN AND MULTI-MINERAL SUPPLEMENT</td>
<td>1266</td>
<td>2.8%</td>
</tr>
<tr>
<td>SINGLE VITAMIN AND SINGLE MINERAL SUPPLEMENT</td>
<td>509</td>
<td>1.1%</td>
</tr>
<tr>
<td>DIETARY SUPPLEMENT OTHER NUTRITIVE SUPPLEMENT</td>
<td>270</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

About 36% of supplements were classified as “dietary supplement – combination/other”. According to the LanguaL\textsuperscript{TM} code, “dietary supplement – combination/other” is a subcategory of “dietary supplement – combination” where 16,531 dietary supplements are included under this category. Among these, 16,522 dietary supplements were classified as “combination/other” subcategory and the remaining 9 products were classified in the “combination” category. We combined these two product types and listed them as one type in the table. However, this is not the only overlap found in this classification system. For herbal and botanical supplements with vitamin or mineral, they could either appear in the “dietary supplement – herbal or botanical” or “botanical supplement with vitamin/mineral” listed in Table 1 since the second category does not specifically exclude the supplements that contain vitamin or mineral. A graphic
presentation of the distribution of each type is shown in Figure 1.

**Normalizing and analyzing ingredient list**

Before normalizing the ingredient information, the distribution of the number of ingredients in each supplement is shown in Figure 1. It can be seen that the histogram peaks at 13,004 for one ingredient, implying about 30% of the supplements have one dietary ingredient. The distribution is also very long-tailed. The 25th percentile, the median, the 75th percentile and the mean for the number of ingredients are 1, 5, 12, and 9.6, respectively. Among 45,455 dietary supplements in the DSLD database, 1,091 items (2.4%) included more than 50 ingredients on their supplement labels, which were not shown in Figure 1.

![Figure 1](image-url)

**Figure 1.** Distribution of the number of ingredients present in the supplements listed in DSLD. Due to the long tail of the data, 1,091 supplement (2.4%) containing more than 50 ingredients were not shown in Figure 1.

These ingredients are classified into 20 categories. The only irrelevant category listed in the database is the “header” category, which only serves the purpose of organizing the dietary ingredient facts, and thus will not be discussed
further. The normalized results for the other 19 categories were shown in Table 2. The “Before Normalization” column lists the number of items extracted from the dietary ingredient fact sheets of all the supplement products in the DSLD, while the “After Normalization” column lists the number of unique items remaining after normalization. Each category was slimmed down by 81.6% to 99.5%, which facilitates faster cross-referencing among different databases.

Table 2. Comparison of the scale of ingredient data in DSLD before and after normalization.

<table>
<thead>
<tr>
<th>Ingredient Category</th>
<th>Before Normalization</th>
<th>After Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blend</td>
<td>842</td>
<td>147</td>
</tr>
<tr>
<td>Bacteria</td>
<td>6535</td>
<td>506</td>
</tr>
<tr>
<td>Protein</td>
<td>5929</td>
<td>145</td>
</tr>
<tr>
<td>Fiber</td>
<td>5395</td>
<td>76</td>
</tr>
<tr>
<td>Element</td>
<td>13160</td>
<td>109</td>
</tr>
<tr>
<td>Chemical</td>
<td>23920</td>
<td>1359</td>
</tr>
<tr>
<td>Vitamin</td>
<td>92604</td>
<td>504</td>
</tr>
<tr>
<td>Fatty Acid</td>
<td>10632</td>
<td>541</td>
</tr>
<tr>
<td>Hormone</td>
<td>512</td>
<td>21</td>
</tr>
<tr>
<td>Animal Part or Source</td>
<td>1896</td>
<td>348</td>
</tr>
<tr>
<td>Fat</td>
<td>20148</td>
<td>139</td>
</tr>
<tr>
<td>Carbohydrate</td>
<td>15919</td>
<td>94</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>25387</td>
<td>739</td>
</tr>
<tr>
<td>Polysaccharide</td>
<td>207</td>
<td>28</td>
</tr>
<tr>
<td>Botanical</td>
<td>80614</td>
<td>6046</td>
</tr>
<tr>
<td>Mineral</td>
<td>54721</td>
<td>484</td>
</tr>
<tr>
<td>Enzyme</td>
<td>9320</td>
<td>341</td>
</tr>
<tr>
<td>Other</td>
<td>34884</td>
<td>4572</td>
</tr>
<tr>
<td>No Label</td>
<td>60005</td>
<td>9213</td>
</tr>
</tbody>
</table>

**Evaluation of supplements term coverage**

The mapping coverage across different terminologies is described in Table 3. The UMLS column presents the number of CUIs mapped by MetaMap. As the mapping process returns explicit semantic types, the quantity of matching CUIs is affected by relevant semantic types as those are not related to dietary supplements were filtered out. But for RxNorm and NDF-RT, the mapping was performed by RxMix that does not output explicit semantic types. RxNorm assigns a unique concept id (RxCUI) for each individual product, making it possible that RxNorm generate more matches than the UMLS.

The mapping coverage across different resources (e.g., online resources, and EHR medication lists) was also compared. As shown in Figure 2, the DSLD database has more unique terms than online resources and the medication list. However, the mapping percentages of UMLS and RxNorm terminologies to the unique terms of DSLD are much lower than the MedList and online database. The NDF-RT terminology produced similar mapping percentage. It is noteworthy that these ingredient categories are not orthogonal. The unique concepts in one category might overlap with the ones in another category; therefore, the final mapping percentage is slightly lower after removing the duplicates.
Table 3. The number of unique concepts mapped into each terminology by the respective ingredient category.

<table>
<thead>
<tr>
<th>Ingredient Category</th>
<th>UMLS (CUI)</th>
<th>RxNorm (RxCUI)</th>
<th>NDF-RT (NUI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino Acid</td>
<td>149</td>
<td>192</td>
<td>188</td>
</tr>
<tr>
<td>Animal Part or Source</td>
<td>59</td>
<td>78</td>
<td>64</td>
</tr>
<tr>
<td>Botanical</td>
<td>1004</td>
<td>1097</td>
<td>515</td>
</tr>
<tr>
<td>Chemical</td>
<td>398</td>
<td>463</td>
<td>454</td>
</tr>
<tr>
<td>Enzyme</td>
<td>75</td>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>Hormone</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Vitamin</td>
<td>128</td>
<td>137</td>
<td>154</td>
</tr>
</tbody>
</table>

Figure 2. Distribution of the number of matched concepts among different resources (DSLD, medication list, and online databases) in UMLS, RxNorm, and NDF-RT terminologies. The percentage of unique terms that have been mapped to each different resource was shown above the corresponding bar.
Although most of the ingredients can be exactly or partially mapped to the UMLS, there is still a portion of ingredients that cannot be mapped to any terminology. Some examples are given in Table 4. Most of these non-mapping results are due to misspelling or errors which occurred during the transfer from physical labels to digital data; instances where the ingredients are simply names created by manufacturer can also lead to a non-mapping result; for botanical ingredients, the name of the ingredients might be coming from another language (e.g., Chinese) that is not in the terminologies; and for chemical ingredients, some nomenclature of organic compounds are not included in the terminologies.

Table 4. Selected examples not covered by terminology.

<table>
<thead>
<tr>
<th>Non-mapping Terms</th>
<th>Problem and Possible Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artichocck</td>
<td>Spelling error - Artichoke</td>
</tr>
<tr>
<td>Sheng Jiang</td>
<td>Names in foreign language – Chinese Pinyin for “raw ginger root”</td>
</tr>
<tr>
<td>Dicyclopentanone</td>
<td>Organic compound nomenclature</td>
</tr>
<tr>
<td>Aminogen</td>
<td>Manufacturer names</td>
</tr>
</tbody>
</table>

Discussion

Due to the known and potential adverse effects on patient safety, DSIs and supplement side effects have attracted a lot of attention due to potential patient risk. Many researchers have found DSIs through pharmacological experiments; however, they usually focus on a small set of supplements or drugs. To largely explore such DSIs from a variety of resources, standardized terminology of supplements is necessary. We have evaluated the terminology coverage among online resources and medication lists in EHR data. Supplement labels contain rich information, including product information and especially complete lists of their ingredients, which are the main components interacting with drugs. Thus a study on the coverage of terminology concerning the supplement ingredients by standard terminologies such as UMLS, RxNorm, and NDF-RT needs to be investigated.

A relatively unique problem for supplements as opposed to comparable over-the-counter and prescription medications is the large numbers of ingredients noted in Figure 2. The use of multiple active agents is fairly infrequent in prescription medications. It shows up more frequently with over-the-counter medications, but it is rare to include more than 3 or 4 components in a product. The supplement data indicate a very large number of ingredients, which may be challenging to practical management of supplement content in secondary document sources such as EHRs. Standardization of ingredient terms is vital for information extraction and information retrieval of dietary supplement in many resources, including both structured data and free texts. Standardization can also promote the accurate use of higher level generic or trade names to represent known ingredient combinations whenever it is feasible. The example of multiple vitamins is a typical case where in the medical record only the multivitamin is noted, since it is known to contain a typical set of usual vitamins. Effective terminology and mappings to higher level names such as the trade name of supplement products can make clinical documentation and research much easier to manage.

The low sensitivity of mapping may also due to the fact that the normalization is unable to cover all possible variations of the ingredients. The rule-based regular expression filters cannot eliminate all possible typos, affixes and duplicates for the ingredients. For example, “copper” and “Cu” were treated as two distinct concepts, but they actually refers to the same concept, which is the copper element. On the other hand, only exact matches were taken into account when mapping the UMLS concept, thus yielding a relatively lower number of mapped concepts. All of these factors contribute to lower mapping sensitivity. Another observation is that RxNorm mapped more ingredient terms in DSLD and medication lists than the UMLS Metathesaurus, which is an anomaly as RxNorm contains more detailed information than UMLS Metathesaurus. For example, there are two distinct RxCUIs that correspond to glycine (1311532 (glycine hydrochloride) and 4919 (glycine)). While MetaMap specifies the semantic type for each CUI, RxMix does not reveal the semantic type difference between the different CUIs, thus making the mapping percentage higher for RxNorm. This could cause further complications as these RxCUIs would be consequently applied in RxMix to find corresponding NUIs. More detailed information is required to eliminate all the one-to-many mappings in order to derive more accurate term coverage for dietary supplements.
This study has certain limitations. We only evaluated the term coverage in a limited number of terminologies, and only considered the exact matches, which may ignore some related concepts from UMLS. This study lays the foundation for additional examination of current drug terminologies and the improved understanding of current dietary supplements. Without terminology examination, further steps such as information extraction would be very challenging. Our future study will investigate how to better represent dietary supplements using existing resources and terminologies to enhance dietary supplement research.

Conclusion

In conclusion, we extracted the product information and the dietary ingredients of all the supplements listed in DSLD. While generic names for the supplements in this database were not available, the dietary ingredient facts became the source of data for analysis of term coverage. The distribution of dietary supplements according to the product types and the ingredient types was derived. The ingredient terms have been mapped to concepts in three different terminologies including UMLS, RxNorm and NDF-RT, to evaluate the ingredient term coverage with the aid of MetaMap and RxMix. It has been observed that the RxNorm provides the best coverage (19.65%).

Acknowledgments

This research was partly supported by the University of Minnesota Grant-In-Aid award (RZ), and University of Minnesota Clinical and Translational Science Institute supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR000114) (Blazar).

References


2061
Clinical Word Sense Disambiguation with Interactive Search and Classification

Yue Wang, MS\textsuperscript{1}, Kai Zheng, PhD\textsuperscript{2}, Hua Xu, PhD\textsuperscript{3}, Qiaozhu Mei, PhD\textsuperscript{1,4}

\textsuperscript{1}Department of EECS, University of Michigan, Ann Arbor, MI, USA; \textsuperscript{2}Department of Informatics, University of California, Irvine, CA, USA; \textsuperscript{3}School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA; \textsuperscript{4}School of Information, University of Michigan, Ann Arbor, MI, USA

Abstract

Resolving word ambiguity in clinical text is critical for many natural language processing applications. Effective word sense disambiguation (WSD) systems rely on training a machine learning based classifier with abundant clinical text that is accurately annotated, the creation of which can be costly and time-consuming. We describe a double-loop interactive machine learning process, named ReQ-ReC (ReQuery-ReClassify), and demonstrate its effectiveness on multiple evaluation corpora. Using ReQ-ReC, a human expert first uses her domain knowledge to include sense-specific contextual words into the ReQuery loops and searches for instances relevant to the senses. Then, in the ReClassify loops, the expert only annotates the most ambiguous instances found by the current WSD model. Even with machine-generated queries only, the framework is comparable with or faster than current active learning methods in building WSD models. The process can be further accelerated when human experts use their domain knowledge to guide the search process.

Introduction

Clinical documents contain many ambiguous terms, the meanings of which can only be determined in the context. For example, the word “malaria” appearing in a clinician note may refer to the disease or the vaccine for the disease; the abbreviation “AB” may mean “abortion,” “blood group in ABO system,” “influenza type A, type B,” or “arterial blood,” depending on the context. Assigning the appropriate meaning (a.k.a., sense) to an ambiguous word, based on hints provided in the surrounding text, is referred to as the task of word sense disambiguation (WSD).\textsuperscript{1,2} WSD is a critical step towards building effective clinical natural language processing (NLP) applications, such as named entity extraction\textsuperscript{3,4} and computer-assisted coding.\textsuperscript{5,6}

Among different approaches to inferring word senses in clinical text, supervised machine learning has shown very promising performance.\textsuperscript{7,8} Supervised machine learning methods typically build a classifier for each ambiguous word, which is trained on instances of these words in real context with their senses annotated, usually by human experts with required domain knowledge. To train an accurate WSD model, a large number of such annotated instances are needed,\textsuperscript{9} the curation of which can be costly as every instance has to be manually reviewed by domain experts. Many methods have been explored in the past to reduce this annotation cost.\textsuperscript{10-14} Among them, active learning, by inviting human experts to directly participate in the machine learning process, has proven to be an effective approach. The premise of active learning is its ability to reduce the number of judgment calls that human experts need to make while achieving the same results as having a fully annotated corpus, thus significantly reducing the amount of human labeling needed.\textsuperscript{14} As such, how to select the most informative instances to present to human experts to annotate is the key to success for the family of active learning based methods.

Existing active learning methods use different strategies to select the most informative instances for annotation.\textsuperscript{15} For example, some select the instance with the least confident prediction or the instance with competing label assignments. However, these strategies suffer from the “cold-start” problem: a number of precisely annotated examples for every sense are usually required to kick off the classifier. Further, a classical active learning procedure does not fully utilize the domain knowledge of human experts. For example, practicing physicians frequently write or read ambiguous words in their notes without any difficulties in conveying or understanding their meaning. They are able to do so largely because of the surrounding context of the ambiguous words; e.g., when AB is used as shorthand for “blood group in ABO system,” physicians know that it commonly appears as “blood type AB,” “AB positive,” or “AB negative.” These contextual words are strong indicators of the sense of an ambiguous word, which is invaluable to a WSD model but remains largely untapped by existing active learning methods.
In this paper, we demonstrate a method that capitalizes on human experts’ domain knowledge to improve the performance of interactive machine learning. We apply a framework that we recently developed, referred to as ReQ-ReC (ReQuery-ReClassify), to the problem of word sense disambiguation in clinical text. Originally designed for high-recall microblog and literature search, ReQ-ReC features a double-loop interactive search and classification procedure that effectively leverages the domain knowledge of human experts. In an outer loop (ReQuery) of the procedure, an expert searches and labels the instances of an ambiguous word along with sense-specific contextual words. Then, a ReQ-ReC system helps the expert compose additional search queries by suggesting other potentially useful contextual words. In an inner loop (ReClassify), the framework requests the expert to annotate the most informative instances selected from those retrieved by all previous queries and then use the results to update the classifier accordingly. An expert can flexibly switch between these two “teaching strategies.” (1) to generate initial examples of a particular sense by launching a keyword search, and (2) to provide fine-grained clarification by labeling the instances selected by the system. Empirical experiments on three different clinical corpora show that this framework is more effective in building accurate WSD models than current active learning methods, even if the expert solely relies on system suggested keywords.

Method

A. The ReQ-ReC Framework

A.1. Sample scenario

To illustrate how ReQ-ReC works, let us consider the following scenario. Suppose we have a set of clinical text snippets (e.g. sentences) all containing the word “AB,” which means either “blood group in ABO system” or “influenza type A, type B.” Our task is to assign the actual sense to each instance. Based on the domain knowledge, a human expert would know that if “AB” co-occurs with the phrase “blood type,” then it likely means “blood group in ABO system;” if it co-occurs with the word “influenza,” then it likely means “influenza type A, type B.” Naturally, the expert would use keywords “blood type AB” to retrieve a set of instances from the text corpus and label them as “blood group in ABO system;” she or he would then search for “influenza AB” and label the retrieved instances accordingly (Figure 1a). These context-sense pairs are used as an initial corpus to warm-start the first round of WSD model learning. The learned model will then be applied to predicting unlabeled instances and ask the expert to further clarify a few boundary cases, e.g. “Labs include influenza AB swab and blood typing.” (Figure 1b). Determining the senses of these boundary cases would allow the model to capture the nuances in language use and quickly improve model accuracy. Later on, the expert may switch between searching for instances and labeling instances. After a few iterations, the expert may start to realize that in phrases such as “AB positive,” “AB” also means “blood group in ABO system.” Through a new search, she or he can quickly label another batch of instances of “AB positive,” which further improves the WSD model (Figure 1c).

Figure 1. An illustrative example of the searching and labeling process of the ambiguous abbreviation “AB.”

From this sample scenario several observations can be made. First, keyword search is a natural interface for domain experts to retrieve cases of ambiguous usage of words and to provide high-yielding, targeted annotation. This process can significantly reduce annotation cost, as human experts are only asked to label instances that are most informative to train the WSD model, while avoiding the need of labeling all instances in a corpus, most of which contribute little to improving the model performance. Additionally, search also benefits the learning algorithm: it provides a warm start in generating an initial model, and subsequent searches further refine the model by covering
other potential senses of an ambiguous word or additional contextual words. Second, while classifying individual instances retrieved by keyword search is necessary for training the model, it is only able to produce a simplistic model, similar to rules. The ReQ-ReC framework therefore asks domain experts to also clarify boundary cases, which informs the model on how to weigh the nuances of language use in clinical text for better sense disambiguation. After being re-trained on these cases, the model becomes more robust and more accurate. In addition, answering these clarification questions might also inspire the human expert to come up with new search queries covering other potential senses of an ambiguous word or additional contextual words that might have not been thought about. Therefore, the two stages – keyword search and active classification – can be used iteratively to inform each other.

A.II. Anatomy of the ReQ-ReC framework

Generalizing from the above example, the double-loop procedure of the ReQ-ReC is illustrated in Figure 2. The procedure operates on an inverted index of the documents so that all keywords, including the ambiguous words and the contextual words, are searchable. The procedure maintains a set of search queries, a pool of retrieved instances, and a WSD model. To start, a human expert first uses her domain knowledge to compose a search query for each known sense, and then the system retrieves an initial set of contexts using the search function. The inner-loop kicks in there, in which the system iteratively presents a small number of instances selected from the current pool of retrieved instances to the expert and asks her/him to assign senses. The WSD model is consequently updated based on the accumulated annotations by the expert, which is then used to reclassify the pool of instances. After a few iterations of the inner-loop, the WSD model’s predictions stabilize on the currently unlabeled instances. At this point, the outer-loop of the system will kick in to recommend new search queries for each sense (the requery process), aiming to retrieve more diverse instances with additional contextual words. These new search queries will be presented to the human expert for review and for further modification. Then, the system will retrieve a new set of instances using the new queries and add them to the existing pool of retrieved instances. After this requery process, the system will start a new inner-loop and continue to update the WSD model. The learning process ends when the expert is satisfied with the predictions made by the WSD model on those unlabeled instances in the newly retrieved pool.

Figure 2. The ReQ-ReC framework.

The framework consists of the following key computational components.

1) Search. The framework uses a standard, Google-like search interface to retrieve instances containing ambiguous words. It can either leverage an existing clinical text search engine or build a just-in-time index over an unlabeled corpus while search is being performed. The search engine’s ranking function can use any retrieval models that take the input of a keyword query and outputs a certain number of instances from the index (using a vector space model,
a boolean retrieval model, or a language modeling approach). After each search, the retrieved instances will be added to a pool, which becomes the basis of the next step, instance selection.

2) Instance selection. In every iteration of the inner-loop, the framework selects a small number (e.g. 5) of unlabeled instances from the current pool of retrieved instances and asks the expert to assign senses. At the beginning of the double-loop procedure, the framework can simply return the top instances ranked by the search engine’s retrieval model. With more and more instances being labeled by the human expert, the system will leverage this knowledge to update the WSD classifier and use active learning strategies to select the next instances for labeling.

3) WSD classification. Given an accumulated set of labeled instances, the WSD classification component learns or updates a multiclass classifier (such as a random forest or a support vector machine) and reclassifies the pool of retrieved instances.

4) Query expansion. When the classifier appears to be achieving a stable prediction on the pool of retrieved instances, the system proceeds to expand the pool in order to cover more contexts in which a sense may appear. This is done through constructing a new query for each sense and retrieving a new set of instances from the target corpus. Query expansion can be done using different methods, such as the Rocchio’s method or semantic term matching. The human expert may either approve this query, edit it, or compose one by her own.

ReQ-ReC is a general framework and each of the key components above can be instantiated in many different ways. In the following subsection, we describe specific implementations of each component.

A.III. Instantiating the ReQ-ReC framework

1) Search. In our current research implementation of the ReQ-ReC framework, we use the Lucene Package to build a search index for each ambiguous word. Instances are tokenized with Lucene’s StandardAnalyzer. To preserve the original form of ambiguous words (“nursing,” “exercises”) and negations (“no,” “without”), we do not perform stemming or stopword removal. We use the Dirichlet prior retrieval function with the parameter \( \mu \) set to 2000, a typical setup in information retrieval literature.

2) WSD classifier. We use logistic regression with linear kernel for WSD classification, implemented by the LIBLINEAR package. If an ambiguous word has two senses, we build a binary classifier; otherwise we build a one-versus-rest multiclass classifier. Logistic regression classifiers output well-calibrated probability predictions \( p(y|x; \theta) \) for each sense \( y \) and each instance \( x \), which will be used by active learning algorithms (\( \theta \) is the classification model parameter). We use presence/absence of the all unigrams appeared in the instance as features. For the L2-regularization hyperparameter \( \mathcal{C} \), we set it to 1.0 across all ambiguous words. This setting is comparable to previous reported studies.

3) Instance selection. In the inner-loop, there are multiple possible methods for selecting the next instance for labeling:

   a) Random Sampling. The algorithm simply selects an instance from the unlabeled pool uniformly at random.

   b) Least Confidence. The algorithm selects the instance \( x \) with the least predicted probability \( p(y^*|x; \theta) \), where \( y^* = \arg\max_y p(y|x; \theta) \) is the most probable sense. Intuitively, the model has little confidence in predicting the sense of instance \( x \) as \( y^* \), therefore it is most uncertain about the sense of \( x \). In this case, expert advice would be needed.

   c) Margin. The algorithm selects the instance \( x \) with the least predicted \( p(y_1|x; \theta) - p(y_2|x; \theta) \), where \( y_1 \) and \( y_2 \) are the most and second most probable senses. Intuitively, the model may not be able to determine if \( y_1 \) or \( y_2 \) is the appropriate sense, therefore it needs further clarification from the human expert.

   d) Entropy. The algorithm selects the instance \( x \) with the highest prediction entropy. High entropy means that the current WSD model considers any sense assignment as almost equally probable. Expert advice is thus needed to resolve the confusion.

In our implementation, we use the margin based active learning strategy to select instances. Note that all four methods can be launched without the search component, which in effect reduces the ReQ-ReC into a classical active learning system. In the evaluation experiments reported in this paper, these methods will be used as baselines for comparison.
4) Query expansion. In the outer-loop, a new query can either be automatically generated by the system and reviewed and improved by human experts, or be composed manually. In this study, we consider the following two extreme strategies: (a) the system automatically generates a new query based on the current status of the WSD model with no human input; and (b) the human expert composes new queries solely based on her or his domain knowledge. These two strategies represent the “worst” scenario and a “desirable” scenario of ReQ-ReC. We use the Rocchio’s method to automatically generate the next query \( q_y \) for every sense \( y \).\(^{18}\) The basic premise of Rocchio’s method is to learn a new query vector that is related to sense \( y \) and far away from other senses.

In fact, we hope that the new query \( q_y \) will not be too close to the known contexts in which sense \( y \) may appear. This would allow the framework to suggest to human experts other contexts of the sense that might not have been thought of. To achieve this goal, we use the “diverse” method developed for high-recall retrieval,\(^{17}\) which generates a new query that balances its relevance to the sense and the amount of diverse information it introduces to the current pool of instances. In the rest of the paper, this strategy is referred to as “machine-generated” queries or the “worst case” of ReQ-ReC.

We also simulate the scenario where human experts use domain knowledge to include contextual words into search queries. To do this, we rank all the contextual words, words appearing in at least one instance of the ambiguous word, by the information gain, i.e. the reduction of uncertainty on the sense of the ambiguous word after seeing a contextual word.\(^{24}\) Top-ranked contextual words are considered as informative and used as search queries to warm-start the initial model learning. In our experiment, the simulated expert guides the first 6 queries using the top 30 contextual words.\(^{†}\) As a simulation of domain knowledge, information gain is computed based on the entire set of labeled instances. Note that information gain is only a crude measure for selecting informative contextual words; human experts can do better with their domain knowledge. This simulation would result in an underestimate of the true performance of ReQ-ReC. We denote this scenario as ReQ-ReC with “simulated expert” queries.

B. Evaluation Methodology

B.I. Evaluation corpora

In this study, we used three biomedical corpora to evaluate the performance of the ReQ-ReC framework.

The MSH corpus contains MEDLINE abstracts automatically annotated using MeSH indexing terms.\(^{8}\) Originally, it has 203 ambiguous words, including 106 abbreviations, 88 words, and 9 terms that are a combination of abbreviations and words. Following previous work,\(^{14}\) we only included ambiguous words that have more than 100 instances so we have sufficient data for training and evaluation. This results in 198 ambiguous words.

The UMN corpus contains 75 ambiguous abbreviations in clinical notes collected by the Fairview Health Services affiliated with the University of Minnesota.\(^{25}\) 500 instances for each abbreviation were randomly sampled from a total of 604,944 clinical notes. Each instance is a paragraph in which the abbreviation appeared. In this study, we excluded unsure and misused senses in training and evaluation.

The VUH corpus contains 25 ambiguous abbreviations that appeared in admission notes at the Vanderbilt University Hospital.\(^{26}\) Similar to the MSH corpus, we only retained 24 abbreviations that have more than 100 instances. Each instance is a sentence in which the abbreviation appeared.

The statistics of the three corpora are summarized in Table 1. We can see that the MSH corpus has the richest context in an instance and the least skewed distribution of senses for an ambiguous word. Because our main goal in this study was to compare the effectiveness of different learning algorithms, we did not further tune the context window size for each corpus.

Table 1. Summary statistics of three evaluation corpora.

<table>
<thead>
<tr>
<th></th>
<th>#Ambiguous words</th>
<th>Average #instances/word</th>
<th>Average #senses/word</th>
<th>Average #tokens/instance</th>
<th>Average percentage of majority sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSH</td>
<td>198</td>
<td>190</td>
<td>2.1</td>
<td>202.84</td>
<td>54.0%</td>
</tr>
<tr>
<td>UMN</td>
<td>75</td>
<td>500</td>
<td>5.4</td>
<td>60.59</td>
<td>73.8%</td>
</tr>
<tr>
<td>VUH</td>
<td>24</td>
<td>194</td>
<td>4.3</td>
<td>18.73</td>
<td>78.3%</td>
</tr>
</tbody>
</table>

\(^{†}\) The first two queries use the top 10 words; the next two queries use the next top 10 words, and so forth.
B.II. Metrics

In this study, we used learning curves to evaluate the cost-benefit performance of different learning algorithms. A learning curve plots the learning performance against the effort required in training the learning algorithm. In our case, learning performance is measured by classification accuracy on a test corpus and effort is measured by the number of instances labeled by human experts. For each ambiguous word, we divided its data into an unlabeled set and a test set. When a learning algorithm is executed over the unlabeled set, a label is revealed only if the learning algorithm asks for it. As more labels are accumulated, the WSD model is continuously updated and its accuracy continuously evaluated on the test set, producing a learning curve. To reduce variation of the curve due to differences between the unlabeled set and the test set, we ran a 10-fold cross validation: 9 folds of the data are used as the unlabeled set and 1 fold used as the test set. The learning curve of the algorithm on the particular ambiguous word is produced by averaging the 10 curves. The aggregated learning curve of the algorithm is obtained by averaging the curves on all ambiguous words in an evaluation corpus.

To cope with the cold start problem of active learning algorithms, we randomly sampled one instance from each sense as the initial training set. To facilitate comparison, we used the same initial training set for random sampling and ReQ-ReC. The batch size of instance labeling was set to 1 for all learning algorithms, so that we could monitor the performance improvement by every increment in the training sample.

To summarize the performance of different learning algorithms using a composite score, we also generated a global ALC (Area under Learning Curve) for each algorithm on each evaluation corpus. This measurement was adopted in the 2010 active learning challenge. The global ALC score was normalized by the area under the best achievable learning curve (constant 1.0 accuracy over all points).

Results

We evaluated six interactive WSD algorithms (one trained on randomly sampled instances, three trained using active learning methods, and two using the worst case and the simulated expert case of ReQ-ReQ) on three biomedical text corpora (MSH, UMN, and VUH).

Table 2 shows the global ALC scores for each learning algorithm on different evaluation corpora. ReQ-ReC with simulated expert queries consistently outperforms all other methods on all three corpora. On the MSH and VUH corpora, even the worst case of ReQ-ReC achieves higher ALC scores than all existing active-learning algorithms. On the UMN corpus, the worst case of ReQ-ReC is slightly outperformed by the margin active learning algorithm. Compared to other active learning methods, the worst case of ReQ-ReC has the highest ALC scores for 164 out of 297 words across three corpora (55.22%) (129/198 in MSH, 20/75 in UMN, and 15/24 in VUH). With simulated expert queries, ReQ-ReC has the highest ALC scores for 206 out of 297 words across the three corpora (69.36%) (156/198 in MSH, 35/75 in UMN, and 15/24 in VUH).

Table 2. Average ALC scores for six learning algorithms.

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Least Confidence</th>
<th>Margin</th>
<th>Entropy</th>
<th>ReQ-ReC worst case</th>
<th>ReQ-ReC expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSH</td>
<td>0.862</td>
<td>0.899</td>
<td>0.900</td>
<td>0.899</td>
<td>0.904</td>
<td>0.913</td>
</tr>
<tr>
<td>UMN</td>
<td>0.854</td>
<td>0.885</td>
<td>0.893</td>
<td>0.878</td>
<td>0.889</td>
<td>0.894</td>
</tr>
<tr>
<td>VUH</td>
<td>0.863</td>
<td>0.871</td>
<td>0.872</td>
<td>0.870</td>
<td>0.878</td>
<td>0.885</td>
</tr>
</tbody>
</table>

Figure 3, 4, and 5 shows the aggregated learning curves of all algorithms on three evaluation corpora, respectively. Results on the MSH corpus present the clearest patterns: the two ReQ-ReC methods learn faster than other algorithms, especially in the beginning stage (first 30 labels). The learning curves of three active learning algorithms are almost identical and much higher than that of random sampling, as previously reported. To achieve 90% accuracy, the best active learning algorithm requires 26 labels on average, while ReQ-ReC with simulated expert queries requires only 17 labels, saving 35% labeling effort.
Patterns on the other two corpora are less significant, due to highly skewed sense distributions. In general, ReQ-ReC with simulated expert queries still achieves the best learning curve than other methods, but with a smaller margin, followed by an active learning algorithm on the UMN corpus and by the worst case of ReQ-ReC on the VUH corpus. Surprisingly, on the VUH corpus, random sampling learns faster than active learning methods at the very beginning. The benefit of active learning kicks in after 20 labels.

**Figure 3.** Aggregated learning curves of 198 ambiguous words in the MSH corpus.

**Figure 4.** Aggregated learning curves of 75 ambiguous words in the UMN corpus.
Figure 5. Aggregated learning curves of 24 ambiguous words in the VUH corpus.

Discussion

The goal of inviting human experts into the machine learning process is to achieve large performance gains with relatively small labeling effort. An active learning process tries to select the next instance such that it brings in as large amount of fresh information as possible for the model to learn from, therefore giving rise to large gains. When asking for the next label, an active learner prefers to ask those instances that represent an unexplored subpopulation and/or instances whose labels the current model is still uncertain about. In contrast, a passive learner randomly picks the next instance from the unlabeled set, regardless of whether it overlaps with a previously labeled one, or whether the model can accurately guess its label, neither of which make the best use of the labeling effort.

WSD model learning benefits considerably from expert queries as a warm start. When the first few queries are informative contextual words, they construct a pool of representative instances. The initial WSD model learned on this representative pool inherits the domain knowledge from the search queries. Human experts can do even better than the simulated expert in composing these queries. Even when the queries are machine-generated, the query expansion procedure also picks up potentially informative contextual words. On the other hand, active learning methods select instances from the entire corpus rather than a representative pool. In the initial learning stage, models are usually poor and their predictions are unreliable. Thus the “uncertain” instances selected by such predictions may not benefit the learning as much as the representative ones. As the model becomes more robust in the later learning stage, the clarification questions raised by active learning will make more sense and labeling these instances can better improve the model.

Different characteristics of text documents affect learning process. In biomedical papers that are formally written (the MSH corpus), an ambiguous abbreviation often appears with its full form for clarification purposes, e.g. “high-risk (HR)” and “heart rate (HR).” The co-occurrence of the abbreviation with its full form greatly makes it easier for both the annotation process and the WSD model. In contrast, an ambiguous abbreviation in clinical notes (the UMN and VUH corpora) is almost never expanded to its full form as abbreviations are typically used to save the time of input. A clinical abbreviation can have many senses that are used in many different contexts. As a result, the annotation process for clinical abbreviations requires extensive search and labeling. Compared to active learning, the ReQ-Rec framework can better assist human experts in building clinical WSD models.

When an ambiguous word has many senses, the sense distribution is often highly skewed: one or two major senses cover more than 90% use cases, while many other senses are rarely used. As we can see in Table 1, word senses of the two clinical corpora are highly skewed (for more than 4 senses, a majority guess has above 70% accuracy).
Skewed sense distribution presents challenge to machine learning. Without abundant labeled instances, it is difficult to learn a WSD model that accurately identifies a rare sense. The classification model will bias towards predicting the major senses and hurt the recall of the rare sense, which becomes an issue for high-stake events such as a rare disease. A straightforward way to cope with the rare sense learning problem is to harvest and label more data for the rare class, for which the first step is to search using contextual words. ReQ-ReC, originally designed for high-recall information retrieval, can be useful in searching for more rare senses.

This study has several limitations. First, in this study we assume the senses of an ambiguous word are known upfront and one instance is already available for each sense, which is a standard setup in the active learning literature. In reality human expert may have knowledge of some but not all of the senses; it is more natural to discover senses on the fly. Second, instead of using the simple bag-of-unigram features, we can use more elaborate features for WSD, e.g. part-of-speech tags, medical concepts (extracted by MetaMap), and word embedding. This could further improve the WSD performance. Third, the framework is only evaluated through simulated experiments and is not evaluated with real users.

Conclusion

In this paper, we describe a novel interactive machine learning framework that leverages interactive search and classification to rapidly build models for word sense disambiguation in clinical text. With this framework, human experts first use keyword search to retrieve relevant contexts in which an ambiguous word may appear to enable targeted, high-yielding annotation. This interactive active learning process, capitalizing on human experts’ domain knowledge, could therefore significantly reduce the annotation cost by avoiding the need to have a fully annotated corpus. Experiments using multiple biomedical text corpora show that the framework delivers comparable or even better performance than current active learning methods, even if human wisdom is not used to aid in the search process (i.e., all search queries are automatically generated by the algorithm). In future work, we will conduct more evaluation studies to assess the performance of the framework using real-world scenarios and real human experts.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work is supported in part by the National Institutes of Health under grant NLM 2R01LM010681-05 and by the National Science Foundation under grant number IIS-1054199.

References

30. He H, Garcia EA. Learning from imbalanced data. Knowledge and Data Engineering, IEEE Transactions on. 2009 Sep;21(9):1263-84.
Representing Residence, Living Situation, and Living Conditions: An Evaluation of Terminologies, Standards, Guidelines, and Measures/Surveys

Tamara J. Winden, MBA1,3, Elizabeth S. Chen, PhD4, Genevieve B. Melton, MD, PhD1,2

1Institute for Health Informatics, 2Dept of Surgery, Univ. of Minnesota, Minneapolis, MN; 3Research Administration, Allina Health, Minneapolis, MN; 4Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

Social determinants of health play an important role in diagnosis, prevention, health outcomes, and quality of life. The objective of this study was to examine existing standards, vocabularies, and terminologies for items related to Residence, Living Situation, and Living Conditions and to synthesize them into model representations. Sources were identified through literature and keyword searches, and an examination of commonly used resources. Each source was systematically analyzed by two reviewers, mapped to topic area(s), and further mapped to a model representation. A total of 27 sources were identified and reviewed. Seven of the sources had no items, i.e. concepts, elements, or values, related to the three topic areas while SNOMED-CT had the most items at 436 followed by the US Census at 174. While none of the identified sources encompassed a complete representation for documenting the three topic areas, their synthesis together results overall in more comprehensive representations.

Introduction

Social and individual behavioral factors play an important role in diagnosis, prevention, health outcomes, and quality of life.1, 2 As defined by the World Health Organization, “social determinants of health are the conditions in which people are born, grow, live, work, and age”.3 Social determinants of health (SDOH) can cause illness, exacerbate chronic illness, but can also prevent disease and improve health. When considered broadly, SDOH includes behavioral components, such as alcohol, drug, and tobacco use; diet; and physical and environmental factors that may influence an individual’s health such as living conditions, social support, occupation and its associated exposures, and physical activities. These factors contribute to mortality as well as account for being a dominant causal mechanism for many types of disease.4–8

With respect to SDOH related to Residence, Living Situation, and Living Conditions, insecurity related to housing, for example, has been associated with poor health among children,9 barriers related to access to health care,10 and the need for chronic disease management.11 Living situation, residence, and physical living conditions all have been shown to have significant impact on a patient’s health outcomes.12, 13 Since SDOH have a significant impact on patient health as risk factors, they should be considered when assessing and prescribing interventions and can serve to influence provider recommendations. For example, if a patient requires a weekly treatment at a distant location, lives in a residential facility, and relies on public transportation, then a weekly treatment plan may not be realistic without additional support, and the patient could become non-compliant. Or, if the patient is homeless, then a treatment plan that requires the use of a refrigerated antibiotic is most likely not a feasible option.

While much work has been done to demonstrate the deleterious effect of behaviors such as alcohol and tobacco use on health outcomes,14 many other social determinants particularly around physical and environmental factors have not been investigated as thoroughly or linked as deliberately to health outcomes. Housing, for example, has been studied with respect to the impact of homelessness on various conditions and housing related exposures.15–24 However, little work has been done around examining the health affects of housing density or with whom the patient lives.25 In a number of cases, the benefits of housing interventions including different types supportive housing have been demonstrated on physical and mental health.26–32 Moreover, knowledge regarding the patient’s physical living space, who the patient lives with, and living conditions would be of benefit clinicians and other stakeholders in providing patient-centered and more appropriate care and services.33

Electronic health record (EHR) systems provide an unprecedented opportunity to collect and analyze patient data at the point of care and for improving our understanding of disease and healthcare outcomes with secondary use of this data. If leveraged fully, SDOH data may be collected at the point of care within these systems and then re-used. For
instance, SDOH can be analyzed along with clinical data to more fully evaluate patient outcomes, generate evidence-based care guidelines, and identify patients who may benefit from special services or interventions or those who may be at higher risk for preventable events. Despite the opportunity of EHRs, standard terminologies and well-designed discrete data collection tools for social history have not been widely developed and incorporated to different EHR systems. In most cases, SDOH documentation may be entered as structured data or unstructured text (e.g., in clinical notes or free-text data collection fields). This leaves a large gap in the data that can be used to enhance patient care as well as facilitate population health research to further the study of SDOH impact to health. The development of standard representations for SDOH information and ultimately optimized data collection tools will ultimately facilitate analytics, clinical decision support, and re-use and interoperability of this information.

To start bridging this gap, work to define and harmonize SDOH standards for inclusion in the EHR has been done,\textsuperscript{34, 35} including recent National Academy of Medicine (NAM; formerly Institute of Medicine) reports recommending social and behavioral domains and measures for EHRs for inclusion in Stage 3 Meaningful Use requirements.\textsuperscript{35} Some work (Chen et al and Melton et al) \textsuperscript{36, 37} has also been done to model social history from clinical notes and public health surveys including residence and living situation. The goal of this study is to expand upon these previous model representations for Residence, Living Situation, and Living Conditions through synthesis of a collection of interface terminologies, standards/specifications, documentation guidelines, and measures/surveys, evaluate the coherence of documentation, and ultimately contribute to a preliminary model representation that will be used to inform design and development of data collection tools in the EHR for these topic areas.\textsuperscript{38}

**Methods**

The topic areas of Residence, Living Situation, and Living Conditions examined in this study are defined and summarized from a literature search in Table 1. This study was performed in three phases: (1) identifying sources potentially containing references to at least one of the three topic areas; (2) analyzing each data source for specific items related to each topic area; and, (3) synthesizing items into comprehensive representations for each of the three target topic areas. For the purpose of this analysis, all references found were weighted equally whether they were a concept, survey question, element, or member of a value list and termed as “items”.

**Table 1: Definitions and sentence/statement examples for target SDOH topic areas.**

<table>
<thead>
<tr>
<th>Topic Area</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RESIDENCE</strong></td>
<td>Describes dwelling types, physical residence, and geographic location. Include safety considerations such as railings or number of floors and steps.</td>
<td>“Apartment building living”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Living in mobile home”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Lives in a nursing home”</td>
</tr>
<tr>
<td><strong>LIVING SITUATION</strong></td>
<td>With whom does the patient live such as roommates, family members, multi-resident dwelling as well as how many others they live with.</td>
<td>“Lives alone”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Total number of people living in the household.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Lives with family”</td>
</tr>
<tr>
<td><strong>LIVING CONDITION</strong></td>
<td>Environmental cleanliness and precautions against infection and disease. Includes sanitation, safety, inadequate water, sewage disposal, heating or cooling, presence of mold, odors, insects, rodents.</td>
<td>“Inadequate heating/cooling”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Presence of lead-based paint”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Cluttered living space”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Presence of mold”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Excessive pets”</td>
</tr>
</tbody>
</table>

Qualifying sources were identified via literature and Internet searches using search terms that included: “<EHR or EMR> documentation”, “social documentation <standards or terminology>”, “social worker documentation <standards or terminology>”, “physical therapy documentation <standards or terminology>”, “occupational therapy documentation <standards or terminology>”, “social determinants documentation”, “<EHR or EMR> residence”, “<EHR or EMR> living situation”, and “<EHR or EMR> living conditions”. Additional sources were suggested by subject matter experts, and professional organization websites for physical therapy, occupational therapy, and social work were examined and also searched using the above list of terms. A source was included in the final set if it contained references to or specific items related to one or more of the three topic areas of Residence, Living Situation, or Living Conditions.

Each of the identified sources were grouped into one of four categories: (1) Interface Terminology, (2) Standards/Specifications/Coding Terminology, Vocabulary, (3) Documentation Guideline, and (4) Measures/Surveys. These categories were partly based upon the sources own definition, as well as the type and function of the source, i.e. how it is used in health IT and patient care. Each individual source was examined to identify specific items related
to each of the three broad topic areas of Residence, Living Situation, or Living Conditions. The source search was limited to assessment-related items and excluded interventional references. Those sources that were available electronically were searched systematically using the provided search tools, such as the IHTSDO SNOMED-CT browser or the standard search mechanisms provided by each source. Sources available in book or paper form were searched manually. The initial list of search terms used was derived from the literature search. The list was further iteratively refined as sources were examined. The final list of search terms used included: “home”, “house”, “housing”, “residence”, “live”, “living”, “lives”, “people”, “mold”, “insect”, “rodent”, “water”, “heat”, “social”, and “density”. All sources were reexamined and searched by two reviewers using this final complete list of search terms. Any discrepancies between the primary and secondary review were subsequently evaluated and the final list of items was amended accordingly to compile a final comprehensive list by consensus. In cases where a source contained items from another standard or terminology source, the duplicates were excluded from the data set in the final analysis.

Each item identified from within each source was then mapped to one of the three topic areas of Residence, Living Situation, and Living Conditions. Each item was then further analyzed and then mapped to specific element axes and values from a previously defined model resulting in an enhanced and more comprehensive model for Residence, Living Situation, and Living Conditions. Element axes and values were added to the previous model representation or expanded on to accommodate findings from this review.

Results

A total of 27 data sources were identified as potentially having applicable items for one or more of the three topic areas. Of the 27 sources, seven sources were excluded during the initial analysis due to lack of detail or specificity. Those included the Meaningful Use Stage 2 Requirements and the EHR Certification Requirements, which did not include specific items directly related to the target topic areas for this study. The Nursing Management Minimum Data Set also did not contain relevant items. The Centers for Medicare Social Work Documentation Guidelines (MSWDG) and the Medicare Rehabilitation Documentation Requirements (MRDR) did contain references to the target topics areas; however, these sources were very high level and did not provide sufficiently detailed information appropriate for this work. Lastly, the Nursing Interventions Classification (NIC) and Uniform Terminology for Occupational Therapy contained items related to interventions and patient activity and not assessments, therefore these two sources were excluded from the final analysis. Secondary review of the 27 sources confirmed these results.

The remaining 20 sources contained applicable items for one or more of the three topic areas. Nineteen of the 20 sources contained some references to Residence totaling 643 items not including the United State Board on Geographic Names, which in itself contained millions of items for national and international named geographic locations (Table 2). SNOMED-CT had the highest number of items related to Residence at 265, followed by the US Census at 160, and third was HL7 at 58 items. Fifteen of the 20 sources contained items related to Living Situation totaling 96 items, the highest being again SNOMED-CT with 62 items, followed by openEHR with 7 items, and the US Census with 6 items. Lastly, 11 of the 20 sources had references to Living Conditions totaling 197 items. The source with the most references to Living Conditions was once again SNOMED-CT with 109 items followed by the Health Indicators Warehouse with 24 items and lastly NANDA-I with 19 items.

Identified items were further manually evaluated and classified using the previously published model representations as a foundation (Table 3). Logical groupings of items were derived and individual items were mapped to an existing element axis or new elements axes were added to accommodate findings related specifically to the three topic areas. The most prevalent mapping was to Residence Detail with 13 sources containing items, and the second most prevalent was Residence Type, which was found in 12 of the 20 sources. Nine sources had items mapping to Living Situation Detail, 9 sources had items mapping to Living Conditions Detail, and 7 had items mapping to Living Condition Type. Of the 20 sources, PhenX Toolkit, openEHR and the US Census had the broadest coverage of elements and values.

Model representations were developed from the mappings. The model representation for Residence uniquely contains items to describe the physical dwelling type, such as “House”, “Apartment”, or “Group Home” as well as details about that physical dwelling including presence and number stairs, and other safety items such as railings. There are also items to collect data about age of the residence, residence build time point as well as geographic location. The model representation for Living Situation includes more detail around Subject since this topic area is related to whom the patient is living with as well as Living situation Detail. Lastly, the model representation for Living Conditions includes an entity to document Living Conditions Type, which refers to the type of hazard such as “mold”, “insects”, or “animals” as well as Living Conditions Detail, which includes details about the Living Conditions Type or more general information about Living Conditions such as “Control environmental pests” (Figure 1).
Table 2: Sources and item counts (*USBGN contains all national and international geographic location names.

**# Items: A=Primary reviewer findings, B=Secondary reviewer confirmation of primary findings, C=Items added by secondary reviewer, i.e. items missed by primary reviewer. ***Total Residence, Living Situation, Living condition counts not mutually exclusive for this source**

<table>
<thead>
<tr>
<th>#</th>
<th>Data Sources</th>
<th># Items**</th>
<th>Total</th>
<th>Total Residence</th>
<th>Total Living Situation</th>
<th>Total Living Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Omaha System: A Key to Practice, Documentation, and Information Management Second Edition</td>
<td>28 27 0 28</td>
<td>11 1</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Nursing Interventions Classifications (NIC)</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Nursing Outcomes Classifications (NOC)</td>
<td>26 24 0 26</td>
<td>19 0</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>North American Nursing Diagnosis Association International Diagnoses: Definitions and Classifications (NANDA-I)</td>
<td>34 22 0 34</td>
<td>14 1</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Clinical Care Classification (CCC)</td>
<td>6 4 0 6</td>
<td>2 2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Nursing Minimum Data Set (NMDS)</td>
<td>6 0 0 6</td>
<td>5 1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Nursing Management Minimum Data Set (NMMDS)</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Outcome and Assessment Information Set (OASIS)</td>
<td>3 3 0 3</td>
<td>0 3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Meaningful Use Stage 2</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2014 Electronic Health Record Certification Requirements</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>IOM Report Social and Behavioral Domains</td>
<td>6 0 0 6</td>
<td>6 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>openEHR***</td>
<td>14 0 3 17</td>
<td>10 7</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)</td>
<td>402 402 34 436</td>
<td>265 62</td>
<td>109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Medical Subject Headings (MeSH)</td>
<td>17 5 1 18</td>
<td>15 2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Logical Observation Identifiers Names &amp; Codes (LOINC®)</td>
<td>4 4 7 11</td>
<td>8 3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Health Level 7 Version 3: Behavioral Health Model</td>
<td>59 37 0 59</td>
<td>58 1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>HL7 Fast Healthcare Interoperability Resources (FHIR)</td>
<td>12 12 0 12</td>
<td>12 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Public Health Information Network (PHIN) Vocabulary</td>
<td>9 1 3 12</td>
<td>10 2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>United States Board on Geographic Names (USBGN)</td>
<td>* * * *</td>
<td>* * * *</td>
<td>* * * *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Occupational Therapy Practice Framework: Domain and Process (3rd Edition) (OT Framework)</td>
<td>4 4 1 5</td>
<td>2 1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Uniform Terminology for Occupational Therapy (UTOT)</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>CMS Social Work Documentation Guidelines (CMSSWDG)</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Amer. Physical Therapy Assoc. Guidelines: Physical Therapy Documentation of Patient/Client Mgmt.</td>
<td>1 1 0 1</td>
<td>1 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Medicare Rehab. Documentation Requirements (MRDR)</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>PhenX Toolkit</td>
<td>32 32 3 35</td>
<td>25 3</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Health Indicators Warehouse (HIW)</td>
<td>34 1 11 45</td>
<td>20 1</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>United States Census</td>
<td>158 26 16 174</td>
<td>160 6</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>855 605 79 934</td>
<td>643 96</td>
<td>197</td>
<td></td>
<td>2075</td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Source item mapping to model elements.

<table>
<thead>
<tr>
<th>Element</th>
<th>Brief Description and Values</th>
<th>OMAHA</th>
<th>NOC</th>
<th>NANDA-I</th>
<th>CIC</th>
<th>NMDS</th>
<th>MA OASIS</th>
<th>IOM</th>
<th>OpenEHR</th>
<th>SNOMED-CT</th>
<th>MeSH Terms</th>
<th>LOINC</th>
<th>HL7</th>
<th>FHIR</th>
<th>PHIN</th>
<th>USBN</th>
<th>OT Framework</th>
<th>APTAG</th>
<th>PhenX Toolkit</th>
<th>HIW</th>
<th>US Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Status</td>
<td>Current, past, or future status</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td></td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>2 Subject</td>
<td>Who (other than patient)</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>3 Patient</td>
<td>For whom the data are being recorded.</td>
<td>■</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>4 Family member</td>
<td>Member of family (spouse, partner,</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>5 Side of family</td>
<td>Maternal or paternal</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>6 Unrelated</td>
<td>Not related to patient</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>7 Other</td>
<td>Adopted</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>8 Negation</td>
<td>Absence</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>9 Certainty</td>
<td>Confidence of statement</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>10 Temporal</td>
<td>Items related to time and dates</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>11 Start date</td>
<td>Date began, exact or estimated</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>12 End date</td>
<td>Date ended, exact or estimated</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>13 Start age</td>
<td>Age began</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>15 Duration</td>
<td>Length of time</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>16 Quantity</td>
<td>Quantity of subject or detail</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>17 Residence Age</td>
<td>Age of dwelling, exact or estimated</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>18 Residence Build Time Point</td>
<td>Point in time when residence was built</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>19 Residence Type</td>
<td>Type of physical dwelling</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>20 Geographic Location</td>
<td>Generic geographic location</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>21 Geographic Location detail</td>
<td>Specific location (country, state, zip )</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>22 Residence Detail</td>
<td>Physical details (levels, stairs, railings)</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>23 Living Situation Detail</td>
<td>Details related to living situation</td>
<td>□</td>
<td></td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>24 Living Conditions Type</td>
<td>Type of sanitation, hazards, clutter,</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>25 Living Conditions Detail</td>
<td>Details related to Living Condition Type</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>✔</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td></td>
<td>□</td>
<td></td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>
**Figure 1:** Preliminary Enhanced Model Representations of Residence, Living Situation, and Living Conditions.
Discussion

While SDOH represent important considerations in the provision of patient care and are also becoming more important in population health management, a number of important gaps in their use continue to exist. Currently, SDOH are not consistently or well-documented in the EHR, particularly the three topic areas on which this work is focused: Residence, Living Situation, and Living Conditions. By working towards synthesis of existing sources to derive enhanced models, we are starting to bridge this gap. Ultimately, the enhanced model representations could be used to inform the design and development of associated EHR documentation tools. To that end, our focus was to err on comprehensiveness in collecting assessment items associated with these topic areas rather than provided or resulting outcomes of those interventions.

A preliminary search for data sources in the form of interface terminologies, standards, specifications, coding terminologies, vocabularies, documentation guidelines, measures, and surveys yielded 27 possible sources that potentially contained items related to the three topic areas. While some sources were eliminated, collating the remaining 20 was highly informative. While no single source was completely comprehensive for all three topic areas, there was overall coverage to support an enhanced model representation. The most broadly comprehensive source was SNOMED-CT, which contained the highest number of relevant items overall, as well as ranking first for the highest number of items for each of the three individual topic areas. The US Census survey ranked second for total items as well as second for Residence (Table 2).

Collectively, the final set of 20 sources were very diverse in specificity primarily because they had distinct but different purposes and uses. For example, OMAHA, NIC, NOC, NANDA-I, and CCC are intended to assist in nursing documentation whereas HL7 is a standard intended for broader data exchange and the US Census and the OASIS are intended to collect data for population surveillance. Lastly, the US Board on Geographic Names was an outlier as it contained millions of very specific items for named geographic locations including city, state, county, country, as well as landmarks and bodies of water. This source could be helpful in standardizing how we model and record geographic locations. As a result of this variability in use and purpose, some sources contained high level, broad concepts, while others contained very specific items and very lengthy, detailed value lists. This diversity presented a challenge in the harmonization phase of the project. For example, SNOMED-CT lists “Housing, local environment and transport finding” with a diverse value list that includes residence types such as “apartment” and “mobile home” as well as living conditions such as “Lives in damp conditions” and “Presence of lead-based paint” while HL7 listed a specific item “primaryResidenceSetting PrimaryResidence” that has an extensive value list associated with it that contained types of physical dwellings. In many cases the lack of specificity caused an item to be counted in more than one topic area. One example of this was found in the Open EHR source that had an item “Description of the home environment”. This could include a description of any of the three topic areas and was counted as such in the final analysis. Ultimately, due to the variability in specificity of the items the data sources, in the final analysis and harmonization, all references to the three topic areas were weighted equally as items whether they were high level broad concepts, questions, elements, or individual values within a value list. This allowed for equal comparison and categorization of the items into the model representation.

Analysis of the final 20 data sources demonstrated that the topic areas of Residence, Living Situation, and Living Conditions are being included in those sources with Residence being most prevalent, specifically Residence Detail (details about the physical structure of the dwelling including safety features) and Residence Type (type of physical dwelling such as house, apartment, group home, nursing home). References to Living Situation were not very detailed. Many of the sources included some indication of whom the patient was living with but the value lists were very non-specific including values such as “family” or “partner/spouse”. Items related to Living Conditions were mostly related to water, sewage, electricity, and heat availability with some references to sanitation specifically insects or rodents. Present, but much less prevalent, were references related to subject other than patient, as were references to temporality, certainty, quantity, and negation, i.e. the absence of an item or condition. Despite these unique challenges, the analysis of these sources, generated 934 items that were ultimately mapped to model elements resulting in enhanced model representations for the three topic areas.

The contributions of this work represent a step towards further informing biomedical standards for the representation of social determinants of health, specifically Residence, Living Situation, and Living Conditions. Next steps will be to further enhance this model representation and ultimately inform the design of EHR data collection tools through the incorporation of EHR unstructured and semi-structured text classifications. One unexpected challenge this work was the complexity of the sources and the difficulty in categorizing them into logical groupings. In future work it may be of interest to examine and develop standard definitions for these types of sources.
Conclusions

While this is a compilation and harmonization of a unique and diverse set of sources, overall, the diversity of the data sources contributed to a broader more detailed model than in previous work. Our results demonstrate that there are many sources that are currently being used to inform data collection with regards to Residence, Living Situation, and Living Conditions. While none of the sources were completely comprehensive, once harmonized they served to inform a more detailed model representation for the three topic areas that can be used towards developing more comprehensive data collection tools in the EHR.

Acknowledgements

This work was supported by the National Library of Medicine of the National Institutes of Health (R01LM011364) and University of Minnesota Clinical and Translational Science Award (8UL1TR000114-02). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References


Evaluating Terminologies to Enable Imaging-Related Decision Rule Sharing

Zihao Yan\textsuperscript{1,6}, Ronilda Lacson, MD, PhD\textsuperscript{1,6}, Ivan Ip, MD, MPH\textsuperscript{1,2,6}, Vladimir Valchinov, PhD\textsuperscript{1,6}, Ali Raja, MD, MBA, MPH\textsuperscript{1,3,6}, David Osterbur, PhD\textsuperscript{5,6}, Ramin Khorasani, MD, MPH\textsuperscript{1,4,6}

\textsuperscript{1}Center for Evidence-Based Imaging, Brigham and Women’s Hospital, Boston, MA, \textsuperscript{2}Department of Medicine, Brigham and Women’s Hospital, MA, \textsuperscript{3}Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, \textsuperscript{4}Department of Radiology, Brigham and Women’s Hospital, MA, \textsuperscript{5}Countway Medical Library, Boston, MA, \textsuperscript{6}Harvard Medical School, Boston, MA

Abstract

\textbf{Purpose:} Clinical decision support tools provide recommendations based on decision rules. A fundamental challenge regarding decision rule-sharing involves inadequate expression using standard terminology. We aimed to evaluate the coverage of three standard terminologies for mapping imaging-related decision rules.

\textbf{Methods:} 50 decision rules, randomly selected from an existing library, were mapped to Systemized Nomenclature of Medicine (SNOMED CT), Radiology Lexicon (RadLex) and International Classification of Disease (ICD-10-CM). Decision rule attributes and values were mapped to unique concepts, obtaining the best possible coverage with the fewest concepts. Manual and automated mapping using Clinical Text Analysis and Knowledge Extraction System (cTAKES) were performed.

\textbf{Results:} Using manual mapping, SNOMED CT provided the greatest concept coverage (83%), compared to RadLex (36%) and ICD-10-CM (8%) (p<0.0001). Combined mapping had 86% concept coverage. Automated mapping achieved 85% mapping coverage vs. 94% with manual mapping (p<0.001).

\textbf{Conclusion:} Although some gaps remain, standard terminologies provide ample coverage for mapping imaging-related evidence.

Introduction

Clinical decision support (CDS) integrated with a computerized physician order entry (CPOE) system can improve workflow efficiency, increase guideline adherence, and reduce the rate of inappropriate imaging in certain outpatient and Emergency Department (ED) settings(1-4). To improve quality of care and reduce waste, the Affordable Care Act was implemented to encourage CDS adoption(1, 5). Additionally, under the Protecting Access to Medicare Act of 2014, health care providers will be required to consult specified appropriate use criteria using a qualified CDS system when ordering advanced imaging for Medicare patients(6).

CDS decision rules are often derived from professional society guidelines, published evidence, and local best practices and are generally described in free text(7). For CDS decision rules to be shareable, they need to be machine interpretable and available in standard representation(8). Recently, a repository of diagnostic imaging decision rules has been developed, which also includes systematic grading of recommendations(7). Although these decision rules are available in semi-structured format, the terminology is not in standard format and limits shareability. Leveraging existing medical terminologies could potentially offer a solution to standardizing concepts within decision rules. In addition, automated approaches for mapping to standard terminologies could enable large scale mapping of decision rules from various CDS systems.

Three terminologies have been utilized previously to retrieve critical imaging findings in radiology - Systemized Nomenclature of Medicine (SNOMED CT), Radiology Lexicon (RadLex) and the International Classification of Diseases, 9\textsuperscript{th} edition (ICD-9-CM)(9). For this study, the newer edition, ICD-10-CM is used. A brief description of these three terminologies follows.

\textbf{SNOMED CT}
SNOMED CT contains more than 311,000 concepts. SNOMED CT is an extensive clinical terminology that was formed by the merger, expansion, and restructuring of SNOMED RT® (Reference Terminology) and the United Kingdom National Health Service Clinical Terms. It is the most comprehensive clinical vocabulary available in English(10).

RadLex

RadLex is developed by the Radiological Society of North America in recognition of limited coverage of radiological concepts by other lexicons(11). RadLex provides a standardized method for indexing radiological concepts in a variety of settings. RadLex consists of approximately 12,000 individual concepts(12).

ICD-10-CM

ICD-10-CM is a clinical modification of the World Health Organization’s ICD-10, which consists of a diagnostic system. ICD-10-CM includes the level of detail needed for morbidity classification and diagnostic specificity. As with ICD-9-CM, ICD-10-CM is maintained by the National Center for Health Statistics. It has more than 68,000 codes, compared to approximately 13,000 in ICD-9-CM(13).

Therefore, the primary goal of this study was to evaluate the coverage of SNOMED CT, RadLex and ICD-10-CM for mapping imaging-related decision rules. As a secondary goal, we assessed automated mapping using Clinical Text Analysis and Knowledge Extraction System (cTAKES), a natural language processing (NLP) tool.

Methods

Source of Decision Rules

This study was exempt from Institutional Review Board review. Imaging-related decision rules were randomly selected from among those in an existing publicly available library of evidence(7). The library currently contains 411 annotated and graded decision rules, derived from practice guidelines and studies published between 1995 and 2014. Specifically, 50 decision rules were selected for this study from five sources – two professional society guidelines (American College of Radiology [ACR] and American College of Physicians [ACP]), local best practice from two healthcare organizations (Ottawa Civic Hospital and Brigham and Women’s Hospital) and a clinical study (Wells Criteria for pulmonary embolism evaluation).

Each decision rule consists of 20 attributes, of which 6 attributes contain values with clinical content that can be expressed using standardized medical terminology. We selected these attributes’ corresponding values for mapping: “imaging modality”, “contrast”, “body region”, diagnosis/symptom”, “clinical logic (if)” and “clinical logic (then)”.

An example of clinical logic – If (Chronic headache) AND (No new features) AND (Normal neurologic examination), THEN MRI of the head without contrast – is included in Figure 1. The other attributes that were not suitable for mapping included names of graders, guideline publishers (e.g. ACR), dates of publications, citations and evidence grades. Full decision rules are publicly available on the library website(14). The list of the 50 rules is included in the supplemental file.

Generating Unique Attribute Values

A total of 300 attribute values were derived from the 50 decision rules (6 attribute values per decision rule). Repeated attribute values were only analyzed once. In addition, stop words, and commonly occurring English words (e.g., with) were removed(15). Attribute values comprised of solely stop words (e.g., “without”) were removed entirely, while those containing partially stop words had only the stop words removed (Figure 1). A total of 75 unique attribute values were generated.
Figure 1. Flow chart of Methods

Manual Mapping

The National Cancer Institute (NCI) Metathesaurus Term Browser, which includes SNOMED CT, RadLex and ICD-10-CM(16), was used to map each unique attribute value to each terminology. To maintain consistency, the “Exact Match” option in the NCI browser was used to search each terminology. We used the UMLS build 2015AB,
which contains SNOMED CT (version, 2013_09_01), Radlex (version, 3_10), and ICD10_2010. Evaluation of match type and computation of coverage are described in the subsequent subsections.

a. Generating unique concepts – A concept is defined as “the fundamental unit of meaning of terms ... and contains all atoms from any source that express that meaning in any way.”(17) If a phrase contains only one word (e.g., Head), then the concept is equivalent to the phrase. If a phrase contains more than one word (e.g., Chronic Headache), then the concept is determined by prioritizing pre-coordinated terms over post-coordinated terms. For example, if the attribute value “chronic headache” results in a simple match (“chronic headache” to “chronic headache”) or a partial match (chronic headache” to “chronic ache”) from any one of the three medical terminologies during mapping, then “chronic headache” would be a concept. Otherwise, “chronic” and “headache” would be two separate concepts that make up the attribute value “chronic headache”. By this definition, any concept with more than one word could be partially matched, while concepts with one word may or may not match. This ensures mapping attribute values to the least number of concepts.

b. Calculating concept coverage – Concept coverage is calculated for each of the three terminologies by calculating the percentage of matched concepts using a single terminology over all concepts from the attribute values. For example, concept coverage for SNOMED CT is defined as the number of unique matched concepts (simple or partial match) using SNOMED CT alone, over the number of existing concepts derived from all attribute values.

c. Categorizing match types – Using methods described by Aronson(18), when mapping attribute value, one of the following could result: “simple match”, “complex match”, “partial match” or “no match”. A simple match results when the concept value is identical to the attribute value (e.g., “pulmonary embolism” to “pulmonary embolism”). A complex match results when individual terms in an attribute value have a simple match to more than one concept, but the entire attribute value does not have a simple match (e.g., “acute ankle injury” to “acute” and “ankle injury”). A partial match results when at least one word of either the mapped result or the attribute value does not participate in the mapping process (“hereditary nonpolyposis colorectal cancer” to “hereditary nonpolyposis colon cancer”), and a no match results when no concept in the attribute value was successfully mapped. All three terminologies were utilized to calculate the highest possible number of matched concepts.

d. Calculating mapping coverage – Further using methods described by Aronson, assuming the attribute value has X words, in which X0 words (X0 ≤ X) participated in the mapping process, and corresponding mapped result has Y words, in which Y0 words (Y0 ≤ Y) participated in the mapping process, the mapping coverage is defined as: 
\[
\frac{2}{3} \left(\frac{Y_0}{Y}\right) + \frac{1}{3} \left(\frac{X_0}{X}\right).
\]
For example, for the attribute value “(Chronic headache) AND (No new features) AND (Normal neurologic examination)” (Fig. 1), there are 7 words in total (“No” is a stop words thus removed). Of the 7 words, “feature” was not matched, while all other words had simple matches. Therefore, X = 7, X0 = 6, Y = 6, Y0 = 6, leading to a mapping coverage of: 
\[
\frac{2}{3} \left(\frac{6}{6}\right) + \frac{1}{3} \left(\frac{6}{7}\right) = \frac{20}{21}.
\]
Abbreviations were un-abbreviated (e.g., MRI would count as three words). Under this definition, simple and complex matches would automatically have a mapping coverage of 1, partial match with a mapping coverage between 0 and 1, while no match would have a mapping coverage of 0. The average mapping coverage is the mean mapping coverage of all 75 attribute values.

Automated Mapping

cTAKES version 3.01(19) was used with YTEX (part of the cTAKES Apache Project) to perform automated mapping. cTAKES was customized with RadLex, the latest releases of the SNOMED CT vocabulary files and ICD-10-CM using the NCI-supported knowledge representation languages, resource description framework (RDF) and the MetamorphoSys’ sub-setting utility, a customization tool provided by the Unified Medical Language System (UMLS) to customize and add source vocabularies to UMLS(20). Custom components were developed to allow cTAKES to take its input from a structured data source and write its output to the YTEX defined schema. cTAKES was applied to all 75 unique attribute values, and the output included concept unique identifiers (CUIs) from the three terminologies(21).

Categorizing match types and calculating mapping coverage were performed similar to manual mapping. We also prioritized pre-coordinated results over post-coordinated results. For example, cTAKES maps “pulmonary embolism”
to the concept “pulmonary embolism” and two other concepts, “pulmonary” and “embolism”. In this case, we would disregard “pulmonary” and “embolism”, and would only consider the simple match result “pulmonary embolism.”

**Data Analysis**

We calculated the concept coverage of SNOMED CT, RadLex and ICD-10-CM individually for all existing concepts, and assessed concept coverage for all three terminologies combined. We used McNemar paired test to compare concept coverage between the three terminologies. We further computed the average mapping coverage for all attribute values resulting from automated mapping, and compared this to the average mapping coverage when we performed manual mapping using paired t-test. A p-value of <0.05 was considered significant.

**Results**

A total of 75 unique attribute values and 220 unique concepts were generated from the randomly selected 50 pieces of evidence.

**Concept Coverage for SNOMED CT, RadLex and ICD-10-CM**

Of the 220 concepts, SNOMED CT provided coverage of 182 concepts, an 83% concept coverage (181/219) rate, significantly greater than RadLex with a 36% concept coverage (79/220, p<0.0001), and ICD-10-CM with an 8% concept coverage (18/220, p<0.0001). When combining all three terminologies, the concept coverage was 86% (190/220). The 8 concepts in addition to the 182 concepts covered by SNOMED CT were contributed by RadLex (Table 1). The unmapped concepts are listed in Table 2.

**Table 1. Concept coverage from each standard terminology**

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Concepts mapped</th>
<th>Concept coverage (Out of 219 concepts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED CT</td>
<td>182</td>
<td>83%</td>
</tr>
<tr>
<td>RadLex</td>
<td>79</td>
<td>36%</td>
</tr>
<tr>
<td>ICD-10-CM</td>
<td>18</td>
<td>8%</td>
</tr>
<tr>
<td>All combined</td>
<td>190</td>
<td>86%</td>
</tr>
</tbody>
</table>

**Table 2. List of 30 unmatched words in any of the three terminologies**

<table>
<thead>
<tr>
<th>BRCA</th>
<th>obvious</th>
<th>Ottawa</th>
<th>hemodynamic</th>
<th>Irrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>nonspecific</td>
<td>noncontributory</td>
<td>rule</td>
<td>Cluster</td>
<td>Critical</td>
</tr>
<tr>
<td>underlying</td>
<td>suspect</td>
<td>exclusion</td>
<td>seen</td>
<td>Ill</td>
</tr>
<tr>
<td>cervicocerebral</td>
<td>labral</td>
<td>criteria</td>
<td>logic</td>
<td>Deterioration</td>
</tr>
<tr>
<td>impact</td>
<td>workup</td>
<td>base</td>
<td>demonstrate</td>
<td>Continued</td>
</tr>
<tr>
<td>dangerous</td>
<td>lifetime</td>
<td>inability</td>
<td>BSGI</td>
<td>image-guided</td>
</tr>
</tbody>
</table>

**Match Categorization and Mapping Coverage: Automated vs. Manual Mapping**

For automated mapping using cTAKES, there were 19 simple matches (25%), 5 complex matches (7%), 49 partial matches (65%), and 2 no matches (3%). For manual mapping, there were 20 simple matches (27%), 34 complex matches (45%), 21 partial matches (28%) and 0 no match (0%). The mapping coverages for manual and automated mapping were 94% and 85%, respectively (p<0.001)

**Discussion**

CDS aids clinicians in decision making, thus reducing medical errors and cost, and promoting more effective care(22). Despite this, most healthcare institutions have limited CDS capabilities(23, 24). One key reason for limited adoption is the predominant use of non-standard approaches to implementing CDS that are often specific to an implementation setting(23, 25, 26). As a result, CDS capabilities developed at one institution may not be easily
transferred to other health care institutions, or even to other types of CDS applications within the same institution(23, 27).

At its core, CDS represents clinical knowledge in a detailed, machine-interpretable format. Machine-interpretable representation, while challenging, is necessary because narrative clinical guidelines often lack the detail and algorithmic specificity required for execution(28). One promising approach to promote CDS capabilities is the availability of machine-interpretable knowledge resources, which can be leveraged across multiple care settings(23, 29-32). However, this approach requires overcoming the heterogeneity that often exists across institutions with regard to patient data and knowledge representation(33). Thus, the adoption of a robust CDS system is critically dependent upon the development and adoption of standards that encompass these facets of CDS delivery.

Various standard terminologies that can represent CDS content are already well-developed(34). Thus, the challenge lies with the fact that multiple terminologies are in concurrent use and the unavailability of content represented using these standard terminologies. As a result, a CDS resource designed for use in one setting may not be readily applicable in another setting. When we assessed concept coverage of SNOMED CT, RadLex and ICD-10-CM, SNOMED CT provided the highest concept coverage at 82.6%. With well over 300,000 concepts, even though not imaging-specific, SNOMED CT provided significantly more coverage than RadLex (35.6%), a radiology-specific lexicon with approximately 13,000 concepts and ICD-10-CM, a terminology primarily for diagnoses and billing with about 68,000 concepts. Although SNOMED CT was able to capture most of the concepts, RadLex presented 8 additional concepts, some of which were specific to radiology (e.g., colonography).

The list of unmapped words (Table 2) covers mostly common English words (e.g., obvious), specific clinical concepts (e.g., hemodynamic), and proper names (e.g., Ottawa). One way to decrease the number of unmapped words is the use of a dynamic suggestion drop down menu when converting narrative clinical guidelines to a structured format. For example, if SNOMED CT is incorporated during the authoring tool when converting guideline recommendations, using synonyms of the unmapped words may prompt authors to use concepts that can be mapped to SNOMED CT.

For the match types of attribute values, manual and automated mapping resulted in similar simple matches, while the majority of attribute values that were mapped as complex matches were mapped as partial matches under automated mapping. The average mapping coverage for manual mapping was significantly higher than that of automated mapping. This result is not surprising as complete matches, defined by combining simple and complex matches (72% in manual mapping; 32% in automated mapping) have a coverage of 100%. Therefore, mapping manually is capable of achieving significantly higher coverage compared to using cTAKES. Some reasons why cTAKES did not have a complete match include deficiency in mapping of abbreviations (e.g., MRCP), adjectives/modifiers (e.g., missing “chronic” in “chronic headache”), and occasionally, incorrect matches (e.g., “without risk factors or neurologic deficit” was mapped to “malnutrition”).

Limitations
This study focused primarily on terminologies that can represent imaging-related guideline knowledge and may not generalize to guidelines in other disciplines. Second, manual effort was required to filter out redundant mapping results from cTAKES. This may have resulted in greater mapping coverage, compared to fully automated mapping. Lastly, while our study addresses the coverage of the three terminologies, it does not taking into consideration the importance of certain words or phrases during the mapping process.

Conclusion
Standard terminologies, such as SNOMED CT, RadLex and ICD-10-CM, provide ample coverage for mapping imaging-related guideline knowledge. Among the three terminologies, SNOMED CT provides the highest amount of coverage. Efforts are underway to further reduce gaps in coverage and increase availability of guideline knowledge, expressed using standard terminologies.

References

2087


Factors Contributing to Dropping-out in an Online Health Community: Static and Longitudinal Analyses

Shaodian Zhang, Noémie Elhadad, PhD
Biomedical Informatics, Columbia University, New York, NY

Abstract

Dropping-out, which refers to when an individual abandons an intervention, is common in Internet-based studies as well as in online health communities. Community facilitators and health researchers are interested in this phenomenon because it usually indicates dissatisfaction towards the community and/or its failure to deliver expected benefits. In this study, we propose a method to identify dropout members from a large public online breast cancer community. We then study quantitatively what longitudinal factors of participation are correlated with dropping-out. Our experimental results suggest that dropout members discuss diagnosis- and treatment-related topics more than other topics. Furthermore, in the time before withdrawing from the community, dropout members tend to initiate more discussions but do not receive adequate response from the other members. We also discuss implications of our results and challenges in dropout-member identification. This study contributes to further understanding community participation and opens up a number of future research questions.

Introduction

Internet and mobile technologies, especially newly emerging social networking applications, are revolutionizing how patients exchange information and social support with care providers, family members/friends, and other patients. Traditionally, patients with life-threatening conditions receive most of the information about their disease from their care providers. While providers tend to focus on the clinical impact of the disease and might ignore the impact of the disease on a patient’s emotional wellbeing and daily life, support groups and more recently online health communities (OHCs), act as a complementary source of support for patients. Previous research suggested that peer patients are able to appreciate each other’s conditions better than health providers, family members and friends, and to exchange necessary emotional support and practical advice of daily health management. Recent years have witnessed increasing popularities of OHCs, with a wide range of formats ranging from discussion forums, to Facebook groups, to dedicated communities like Patientslikeme. Critical to studying OHCs’ impact on their members is characterizing and understanding the patterns of participation in a community. Researchers have studied whether users actively participate or lurk, as well as when they decide to withdraw from the community permanently. Lurking—the phenomenon of users browsing the content but not actively participating in discussions—has been shown to correlate with lower perceived social support and diminished emotional benefits when compared to active participation in a community. Dropping-out—i.e., stopping participation or leaving an intervention, such as an online community, altogether—when examined across members indicates the level of activity in an OHC. For instance, Eysenbach and colleagues reported that the phenomenon of attrition (or dropout) is particularly common in online-based interventions, with more than 90% of study subjects quitting throughout Internet-based studies. In the case of OHCs, understanding factors associated with dropping-out might help identify opportunities for more targeted support of members, and more generally identify for which members participation in an OHC is beneficial and for which it is not. Wang and colleagues examined how type of information received affect users’ choices between staying and leaving, and suggested that informational support is positively correlated with dropping-out while emotional support is positively correlated with staying active in the community. Zhang suggested that information and small group interactions, like emotions, also play a key role in retaining users. Sadeque and colleagues proposed a supervised model to predict dropping-out, and found that factors like time since last activity were predictive. To date, however, it is still unclear which other factors of individual members are moderating dropping-out from online health communities, such as topic of discussions, users’ sentiment expressions, and interactions among users.

Previously we carried out longitudinal analysis of members’ posting history in an online health community to explore how participation affects members’ sentiment, and how different factors impact sentiment change through time.
found that members show an increasingly positive sentiment at the early phases of participation, and that later changes may be correlated with different variables like age and cancer stage. In this study we carry out a series of static and longitudinal analyses, which take topic of discussions, sentiment, and user interactions as variables of interest. We explore if and how these factors correlate with members’ decisions of dropping-out. Because there is no explicit marker for any participant to convey they dropped out of the community, we explore different approaches to determining that a member dropped out. To explore factors in context of dropping-out, we leverage established machine-learning-based methods for sentiment analysis and topic classification of a given member’s posts. We hypothesize that dropout members discuss more disease-specific topics, express more negative sentiments, and interact with other members less actively than the members who stay active in the community. Furthermore, we hypothesize that characteristics of dropping-out can be detected by investigating patterns of changes of these factors.

Methods

The basic workflow of our analysis is as follows. First, we identify members that have dropped out from the community, i.e., members who had history of active participation in community, but have been inactive for a certain amount of time. We collect a set of variables for each members throughout their history of participation in the community. We compare distributions of each variables between dropout members and other members. The longitudinal analyses focus on dropout members to investigate if any patterns of changes of variables exist before they drop out the community, with respect to their sentiment expressions, topics of discussions, and interactions with other members.

Data set

The data set we relied on is from the publicly available discussion board of breastcancer.org, one of the most popular online breast cancer communities. The discussion board is organized in several forums, each with threads and posts. At the time of data collection (January 2015), the dataset consisted of 3,282,008 posts in 121,474 threads, published by 57,424 community members. We extracted post content, public available meta-data like author ID and author name, as well as post signatures including user profiles of self-reported demographics, diagnosis and treatment histories.

Identifying dropout members

Identifying which members in a public community dropped out is not a trivial task. In practice, it is impossible to determine with absolute certainty a dropout member from a public community solely based on changes of posting activity, since an inactive user can always return to the community and resume participation. Moreover, in many communities, like our community of interest in this study, there is no publicly available information about members and their login patterns; and as such the only available information relates to their posting activity. Thus, a member could withdraw from posting content, but still act as a lurker.

To identify the cohort of dropout members for our study, we explored different heuristics. We defined a user in the breast cancer forum as a dropout member, if she has posted more than \( n \) times in the community (i.e. had some history of posting activity), but has been inactive for at least \( t \) years at the time of data collection. The first cut-off is to ensure that users we identify are users who participated in the community discussion meaningfully, instead of one-time information seekers or users who just chimed in a limited number of discussions without real information or support exchanges with other members. The second threshold is to exclude members that may return to the community in the near future, as we assume that users who have been inactive for longer time are less likely to return.

In this particular study, \( n \) and \( t \) were experimentally set as 10 posts and 3 years. As such, for the remaining part of this paper, dropout members refer to those who have posted more than 10 times in the community, and whose most recent post was before January 2012 (three years before January 2015).

Longitudinal analysis of dropout members

Three specific variables were studied to examine if they are correlated with dropout: topic of discussion, interaction with other members, and sentiment expression. These three variables are important building blocks of OHC content and member characteristics, and have been investigated in a wide range of previous studies.

Our research hypotheses are as follows:

2091
Table 1. Topics considered for analyzing breast cancer forum posts.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative</td>
<td>ALTR</td>
<td>alternative, complementary, and integrative medicine</td>
</tr>
<tr>
<td>Daily</td>
<td>DAIL</td>
<td>daily cancer-related experiences</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>DIAG</td>
<td>diagnoses, measurements, and results of tests</td>
</tr>
<tr>
<td>Finding</td>
<td>FIND</td>
<td>health findings, signs, symptoms, and side effects</td>
</tr>
<tr>
<td>Health Systems</td>
<td>HSYS</td>
<td>health systems patients interact with, e.g., nurses, doctors, practices, hospitals, and insurance companies</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>MISC</td>
<td>greetings, and uninformative text, or which does not fit under any other label</td>
</tr>
<tr>
<td>Nutrition</td>
<td>NUTR</td>
<td>nutrition</td>
</tr>
<tr>
<td>Personal</td>
<td>PERS</td>
<td>personal anecdotes, information</td>
</tr>
<tr>
<td>Resources</td>
<td>RSRC</td>
<td>links, pointers, and quotes towards informational resources</td>
</tr>
<tr>
<td>Test</td>
<td>TEST</td>
<td>testing procedures (but not results of tests)</td>
</tr>
<tr>
<td>Treatment</td>
<td>TREA</td>
<td>treatments, including procedures, medications and therapeutic devices</td>
</tr>
</tbody>
</table>

1. Dropout members are more likely to discuss certain topics such as cancer treatments and their side effects, and show certain patterns in topic transitions, before they drop out. These topics and topic transitions may indicate end of cancer treatment journeys, which are usually followed by participation withdrawal.

2. Dropout members receive inadequate social support from other members. They ask questions and seek support more often than other members, but receive less responses. These may indicate lower levels of social support reception leading to decreased sense of belonging, a phenomenon known to be vital to self-perceived effectiveness of community usage.\(^{30}\)

3. Dropout members express more negative sentiment in general, or in their final stage of participation, which indicates a declining level of satisfaction towards community participation.

**Dropout and topics.** To investigate whether topics of discussions correlate with dropping-out, topics of posts must be identified. In this study, topics of posts were identified using a supervised machine-learning tool based on convolutional neural networks (CNN).\(^{31}\) The tool was trained on an annotated data set consisting of 9,016 posts, which was sampled from the same data set used in this study. Eleven topics were considered, which are relevant to describing the information needs of the breast cancer community members. The topics range from disease diagnosis, treatment, to more personal issues like daily lives and nutrition; Table 1 lists all topics with their descriptions. Of note, each post can be annotated with several topics (and as such, the topic identification task is cast as a multi-label classification). Overall, the classifier can identify topics of discussions with around 65% F score across all 11 topics. Further methodological details and detailed system evaluation of the CNN classifier can be found in the cited paper.\(^{31}\)

To characterize the impact of discussed topics, for each user (either dropout or non-dropout members), we aggregate numbers of topics of all posts authored by the user, and average the topic numbers by the total number of the user’s posts. As such, an eleven-dimensional distribution of topics can be established for a member in the forum, representing frequencies of topics discussed by the user.

Armed with distributions of topics for all users in the community users, we first did a multivariate t-test to examine the difference of topic distributions between posts of dropout members and posts of other members. For each topic, we then carried out a univariate t-test, adjusted by Bonferroni correction due to multiple comparisons, between the dropout members and other members in the community to test if a significant difference exists. These two static analyses identify the distributional differences between topics of discussions between dropout members and other members.

Finally, we examined how the averaged frequencies of topics change through time for dropout members before they actually quit the community from a longitudinal standpoint, to investigate whether certain patterns of changes could be detected.
**Dropout and interaction.** Member interaction is the primary medium of exchanging social support, which can be complex in online health communities. In this study, we considered two basic aspects of user interactions: number of initial posts versus number of reply posts, and average number of responses received from other members in the community. Initial posts are those posts initializing threads of discussions, which are usually question asking or help seeking which represent needs of support requesting. Previous research has reported that initial posts are vital part of interactions amongst members, and are usually more negative emotionally. Reply posts, usually representing support providing, are those posts responding to the initial posts, which can exert positive influence on the discussion originator (i.e., author of the initial post). As such, the ratio of number of initial posts to the number of reply posts can be seen as how often the user seek support from others rather than actively provide support to others. Average number of responses received when initializing discussions, on the other hand, represent how much social support in average members receive from other ones. Previous studies have suggested that support providing and receiving may have different effects on perceived benefits.

For each member, we counted the number of their initial posts, the number of their reply posts to other member’s threads, and the number of responses received from other people when initiating a thread. We then calculated the two measures described above, and examined how these numbers differ between dropout members and other members. We relied on a Chi-square test (for initial vs. reply) and t test (for number of replies). Like for the topics, we also examined how these numbers change longitudinally before members’ dropping-out.

**Dropout and sentiment.** Sentiment expression reveals how positive the author’s emotion is when posting. We rely on a supervised classifier for sentiment analysis, which is described in. For each of the post, a sentiment score was calculated, representing the degree of positiveness of the overall sentiment expression. Based on the sentiment scores, we first identified if a significant difference exists between the averaged sentiment scores of posts published by dropout members and posts published by other members, by doing a t-test. Second, we illustrated how sentiment of posts changed through time as dropout members approached the time point when they withdrawn from the community, to see if a decline of sentiment actually happened as suggested by our hypothesis.

**Results**

**Identifying dropout members**

6,338 dropout members were identified using our definition, corresponding roughly to 11% of all users that have posting history in the breast cancer forum. When accounting for all users who have posted more than 10 times (i.e., “meaningfully active”) in the community, the dropout members amounted to 42% of these 15,199 users. The identified dropout members posted 570,932 posts in total in the breast cancer forum, with each one posting 90.1 posts in average. The average posting number is roughly the same as the average across all users posted more than 10 times (91.8). 195 out of these 6,338 dropout members have been highly active in the forum, with each of them posted more than 500 times. These “super-users”, although relatively small in number, contributed to roughly 45% of posts identified.

**Longitudinal analysis of dropout members**

**Dropout and topics.** The multivariate t-test between the topic distributions of posts contributed by dropout members and other members respectively yielded a result which supports a difference with p-value less than $10^{-16}$. Average prevalence of each topic for the two types of members is given in table 2 with corresponding p-values based on the univariate t-tests. We did not include MISC in the table because it is a default topic category only given to those posts which are not assigned any topics otherwise. We used 0.001 as the threshold of p-value for significance. Five topics amongst all ten show significant differences in average numbers between dropout members and other members. Specifically, dropout members posted more relevant to diagnosis and treatment, but less about nutrition and daily matters. The hypothesis that dropout members discuss more about treatment and diagnosis than other members is thus supported.

Figure 1 shows how topic frequencies change through time as dropout members approach the time point of withdrawing. The way we illustrate the changes is as follows. For each topic category, we plotted change of its average frequencies in all posts that were published a certain length of time before their authors’ respective dropout time. We used week, days, and post orders as three different measures to show both long term and short term effects. For exami
Table 2. Average prevalence of topics (per post) in posts of dropout members and other members. P-values are calculated by a t tests adjusted by Bonferroni correction. We use 0.001 as the threshold of p-value for significance.

<table>
<thead>
<tr>
<th></th>
<th>ALTR</th>
<th>DAIL</th>
<th>DIAG</th>
<th>FIND</th>
<th>HSYS</th>
<th>NUTR</th>
<th>PERS</th>
<th>RSRC</th>
<th>TEST</th>
<th>TREA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropout</td>
<td>0.002</td>
<td>0.059</td>
<td>0.099</td>
<td>0.063</td>
<td>0.081</td>
<td>0.034</td>
<td>0.274</td>
<td>0.013</td>
<td>0.009</td>
<td>0.053</td>
</tr>
<tr>
<td>Others</td>
<td>0.002</td>
<td>0.074</td>
<td>0.093</td>
<td>0.063</td>
<td>0.078</td>
<td>0.039</td>
<td>0.279</td>
<td>0.017</td>
<td>0.010</td>
<td>0.046</td>
</tr>
<tr>
<td>p-value</td>
<td>0.226</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
<td>0.953</td>
<td>0.030</td>
<td>&lt;0.001*</td>
<td>0.162</td>
<td>&lt;0.001*</td>
<td>0.247</td>
<td>&lt;0.001*</td>
</tr>
</tbody>
</table>

ple, a point $(1, 0.3)$ in Figure 2(a) or 2(d) represents that the average frequency of the corresponding topic of all posts that are published in the final week of their authors’ participation is 0.3. Except for an trend for a higher frequencies of DIAG and HSYS posts in the final weeks, no significant changes of topic frequencies were identified before members’ dropping-out.

Figure 1. How topic frequencies change through time before members’ dropping-out. X axes, which are in reserve order, represent the time point before members’ dropping-out. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

Dropout and user interaction. 121,193(3.9%) of all posts in the forum are initial posts of threads. Among them, 31,277 were posted by dropout members, which are 5.5% of all dropout member publications. However, the Chi-squared test indicates no significant difference between dropout members and other members in terms of ratio of initial to reply posts, with a p-value over 0.9. Across the entire forum, an initial post can receive 24.4 replies in average. Dropout members, in particular, can receive an average number of 23.7 replies throughout their community engagement when initializing discussions. A t-test between the numbers of dropout members and other members indicate no significance with p value 0.69. As such, the hypotheses that dropout members receive less reply from other people and that post initial posts more often in the community are both rejected.

In contrast, the ratio of initial posts increases towards dropout time (Figure 3). It is particularly significant from a longer term standpoint, where the ratio of initial posts dramatically increase from around 5% to over 10% in the last 10 weeks of participation before dropping-out. We carried out a supplementary t-test, in which we compare all posts in final 10 weeks and posts before 10 weeks in terms of the initial/reply ratio, and indeed found a significant difference between the two with p value less than 0.001. Short term changes can also be observed, particularly in the final 5 days. Meanwhile, in term of number of replies received, a landslide can be observed in the week view, which roughly
accompanies temporally the ratio increase of initial posts.

Figure 2. How percentage of initial posts and number of replies change through time before members’ dropping-out. X axes, which are in reserve order, represent the time point before members’ dropping-out. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

Dropout and sentiment. The average sentiment score (probability of being positive) for all posts in the community is 0.786, while the average sentiment score of dropout member authoring is 0.788, with no significant difference according to a statistical t-test. Longitudinally, an insignificant decline of sentiment can be observed from the week view, but no other patterns can be found. Although we found a tendency of posting more initial posts in the final stage of participation in the previous analysis, no patterns of sentiment change is visible when initial posts and reply posts are considered separately. In contrast to our expectation, dropout members not necessarily express more negative emotion in discussion, and no significant changes of sentiment can be detected before they drop out.

Discussion

Principle findings

Our first hypothesis that dropout members are more likely to discuss certain topics is supported by our experimental results. We find that dropout members tend to discuss more about disease diagnosis and treatment, but less about daily issues and nutrition. Topics of treatment and diagnosis are common in posts that tell stories of one’s cancer journey, or that describe cancer treatment experience. On the contrary, more daily-matter issues like exercises and nutrition are less focused by these users. Not many significant patterns of topic changes are identified longitudinally, except increased frequencies of health system and diagnosis in the final weeks before dropping out. This seems to suggest that although dropout members are more interested in certain topics in general, they do not necessarily shift their focus drastically throughout their participation. The increasing frequency of DIAG is interesting, however. One possible explanation may be that many dropout members were patients who were diagnosed with cancer recurrences or metastasis, which may be followed by the deterioration of the disease.

Our second hypothesis, with respect to user interactions with other members, is partially supported by the results. We originally expected that dropout members receive less replies from other members, which represents a lower level of social support received from other users, and that dropout members post initial posts more often, which represents that they are more likely to be information seekers rather than social support providers. Previous research in online social support groups suggested that emotional support providing is an important motivation of participation and is beneficial to the providers themselves socially, which is a factor that are expected to be negatively correlated with
Figure 3. How average sentiment score changes through time before members’ dropping-out. X axes, which are in reserve order, represent the time point before members’ dropping-out. The first three figures show the average score of posts including both initial and reply, and the last three figures distinguish the two. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

attrition.

However, in our static analyses, no significant differences are identified in the static analysis between dropout members and other members with respect to number of replies received, or ratio of initial posts to number of reply posts. The result may have two possible explanations. The first is that neither of the two measures can truly represent the degree of social support exchange in online health communities, and the other is that OHC users, particularly BC forum users, are different from online social support group members studied in previous research in how they perceive and understand benefits.

Although static comparison finds no difference, longitudinally we indeed find a rather significant increased ratio of initial posts at the end of user participation, as well as an insignificant drop of numbers of received replies, which is consistent with findings in the previous research that number of replies is important predictor of dropout. The change is particularly dramatic in the final few weeks from the week view, and in the final 5 days from the day view.

This result, along with results from the static analysis of interactions as well as from previous analyses of topics, possibly shows a more complete picture of dropping-out: dropout members, in terms of support seeking and support providing, are identical to other community members in most of the times throughout their participations; however, certain events, which may be from the real lives of the users such as recurrence of cancer, trigger online behavioral changes and make the users seek much more support than before. At this moment, if these members don’t receive adequate support, dropout may eventually happen.

Our final hypothesis that users express increasingly negative emotions in posts are not supported by our analysis. No significant difference is found between dropout members and other members, and no clear patterns can be identified longitudinally. The results contradict findings in previous research that usages of emotional keywords are associated with dropping out, possibly because keywords of emotions cannot truly represent sentiment. Synthesizing the sentiment and interaction results seems to suggest that changes at the end of participation are mostly peaceful in sentiment, with no evident clue emotionally.
Table 3. Number of dropout members identified as the cut-off t changes.

<table>
<thead>
<tr>
<th>t cut-off</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td># of dropout members</td>
<td>13,997</td>
<td>9,677</td>
<td>6,338</td>
<td>3,864</td>
<td>2,311</td>
<td>925</td>
<td>210</td>
<td>76</td>
<td>32</td>
<td>11</td>
</tr>
</tbody>
</table>

Validity of our dropout member identification method

In this study, we rely on a straightforward method to identify more than 50 thousand dropout users, assuming that these users having an active history but having been inactive for at least 3 years. For a public online health community like breast cancer forum in which the community has no control over its members, identifying dropout members with absolute accuracy is infeasible. Because of the unavailability of gold-standard, evaluation of such identification method is challenging.

In our method, the most tricky part is to choose the \( t \) cut-off, which represents the minimal length of inactiveness for a member to be considered as dropped out. A larger \( t \) would definitely bring a set of dropout members with higher precision, but may excludes eligible dropouts incorrectly. Given the fact that most members joined the community in recent years and the forum was getting increasingly popular, a large \( t \) would lead to a small sample size. As such, the problem becomes a precision-recall trade-off, and our task is to finds the best value that balance the two properly. The oldest posts of our data set date back to Sep 2004, which is roughly 10 years before the data collection. To see how the cut-off impacts the sample size, we show in table 3 number of dropout members identified by setting \( t \) from 1 to 10. It can be seen that sizes of samples shrink rapidly when larger \( t \) is used.

The major false-positives of our method are the users that return to the community after long time inactiveness. To quantify the prevalence of these comebacks, we designed a sanity check experiment in which we calculate the percentage of users who have been inactive for more than 3 years in the community, anytime in the history, but return to the community after the long break, over the total number of users who have been active for more than 3 years. The number we get is 1.2%, which suggests a relatively good precision of our identification method.

What we learned from our topic analysis that dropout members focus more on diagnosis and treatment related themes also reminds us that users may drop out of the community because of death. Their escalated interest in diagnosis and treatment related issues may just be a signal of cancer metastasis, or unsuccessful treatments which may be followed by deterioration of the disease. These members leave the community not because of dissatisfaction towards community usage, and should usually be excluded in the attrition analysis. Similar to the issue of returning of inactive users, in public online communities there is no way to accurately identify dead members. To investigate how much this confounder impacts our results, we extract cancer stage information from user signatures, exclude cancer stage IV users, and replicate all analyses. The rationale is that stage IV users are the ones most likely to leave the community because of death, while stage 0 to stage III breast cancer are believed to have quite high 5-year survival rate. These supplementary analyses show identical findings as we demonstrated previously, and the exclusion of stage IV users does not impact the results. It is noteworthy, however, that the result does not indicate the nonexistence of impact of dead members on our study since not all users have accurate profile information in signatures.

Limitations and Future Work

There are several limitations of this work, which will be important parts of our future work. First, methodologically, our methods identifying dropout members, topic of discussion, and sentiment of posts are all based on automated algorithms that are not 100% accurate. Manual analysis over a small sample set might be a good way to complement the automated analyses. Efforts will also be made in creating better computational tools for topic and sentiment analysis in future work. Second, in this paper we only consider three variables of participation, while many other important ones are not covered such as members’ age, social status, whether there are community debates or conflicts, etc. These variables of interest will be the focus of our future work to study how they impact users’ decision makings. Third, our analyses, especially static ones, present only statistical differences, and such differences do not guarantee real clinical or psychological differences. In this paper, we do not try to identify causations and provide explanations; rather, we focus on detecting interesting correlational patterns that worth exploring by future research with rigorous experimental study designs. Our analysis in this study has captured some signals of what factors may be contributing to
dropout, and we have proposed that it is likely that real life events may trigger the change the way members exchange social support and their foci of discussions. We believe that our study is a showcase of how quantitative methods can be used to analyze OHC content at scale for hypothesis discovery. Further analysis, both qualitative and experimental ones, can be carried out to examine these hypotheses. Forth, a better distinguish between dropout member and lurking members should be made more systematically. Identification of dropout members in public online health communities is a challenging task, and should worth more exploring in the future. For example, one possible solution might be doing expert annotations, followed by supervised learning models. Finally, this study was conducted on a single online health community. It will be interesting to see the impact of these factors in different communities specific to breast cancer as well as to other conditions.

Conclusion

This paper presents a quantitative exploratory study over a popular and active public online breast cancer community to identify the characteristics of members that quit participation. We investigate correlations between community dropping-out and different factors: topic of discussion, sentiment of post, and user interactions in the community. When conducting such a quantitative study, one important methodological question pertains to identifying at scale the users who drop out. We explore strategies to identify such members, as well as static and longitudinal analyses of members’ post history. Dropout members did not show any significantly different patterns in sentiment change when compared to other members in the community. Our findings suggest however that dropout members (1) tend to focus more on diagnosis- and treatment-related topics; and (2) exhibit increased needs of social support at the ending phase of participation, which are less and less fulfilled by other members.

Acknowledgments

This work is supported by a National Science Foundation award (#0941339) and a National Institute of General Medical Sciences award (R01 GM114355).

References

[16] Frost J, Massagli M. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another’s data. J Med Internet Res. 2008;10(3):e15.
DCDS: A Real-time Data Capture and Personalized Decision Support System for Heart Failure Patients in Skilled Nursing Facilities

Wei Zhu<sup>1,2</sup>, Lingyun Luo<sup>2</sup>, Tarun Jain<sup>2</sup>, Rebecca S. Boxer<sup>3</sup> MD MS, Licong Cui<sup>1</sup>, PhD, Guo-Qiang Zhang<sup>1</sup>, PhD

<sup>1</sup>Institute of Biomedical Informatics, University of Kentucky, Lexington, KY
<sup>2</sup>Department of EECS, Case Western Reserve University, Cleveland, OH
<sup>3</sup>Divisions of Geriatrics and Cardiology, University of Colorado, Aurora, CO

Abstract.

Heart disease is the leading cause of death in the United States. Heart failure disease management can improve health outcomes for elderly community dwelling patients with heart failure. This paper describes DCDS, a real-time data capture and personalized decision support system for a Randomized Controlled Trial Investigating the Effect of a Heart Failure Disease Management Program (HF-DMP) in Skilled Nursing Facilities (SNF). SNF is a study funded by the NIH National Heart, Lung, and Blood Institute (NHLBI). The HF-DMP involves proactive weekly monitoring, evaluation, and management, following National HF Guidelines. DCDS collects a wide variety of data including 7 elements considered standard of care for patients with heart failure: documentation of left ventricular function, tracking of weight and symptoms, medication titration, discharge instructions, 7 day follow up appointment post SNF discharge and patient education. We present the design and implementation of DCDS and describe our preliminary testing results.

Introduction

Heart failure (HF) is an important healthcare issue because of its high prevalence, mortality, morbidity, and cost of care. With the aging of the population, the impact of HF is expected to increase substantially, due to the fact that more older Americans are hospitalized for HF than for any other medical condition. More than 5 million people in the United States have heart failure. One in 9 deaths in 2009 included heart failure as a contributing cause. About half of people who develop heart failure die within 5 years of diagnosis. Overall, HF costs the nation an estimated 32 billion each year [2, 3, 4]. The trial of Heart Failure Disease Management Program in Skilled Nursing Facilities (SNF Study) is a study funded by the National Heart, Lung, and Blood Institute (NHLBI) [5] to evaluate if management of HF in SNFs could improve patient-centered outcomes and reduce hospitalization. A wide variety of data relevant to patient disease management is collected in this study. Thus, it calls for a data management system which accommodates a variety of data types as well as unique data capture needs, such as complicated medication entry. Access to clinical data is another key part in making a personalized healthcare strategy and enhancing the operation and adjustment of HF-DMP. To support real-time healthcare decision, a set of data analysis functionalities such as validation, retrieval, and visualization is required.

An economical, flexible, and scalable data entry system for capturing patient information which supports interoperation plays an important role in large scale, multi-center clinical trials. Unlike paper- or document-based forms, rapidly-deployed web-based electronic interfaces have the advantage of ensuring data quality through built-in data validation mechanisms. In addition, a well-designed electronic system can provide intuitive data visualization and better support health care decision making. While REDCap [1] (Research Electronic Data Capture) is a well-known data capture tool which meets a large percentage of SNF Study needs, there are unique features that REDCap does not provide.

To address this, we present a Real-time Data Capture and Personalized Decision Support System (DCDS) for SNF Study, focusing on real-time data capture and personalized decision support. DCDS supports automated score calculation, visualization of symptoms, and medication history retrieval. Patients benefit from the system because intuitive visualization of symptoms by caregivers may lead to enhanced care, improved outcomes, and reduced risk of death. The caregivers benefit from the system because it is easier for them to make sense of voluminous data and make prompt clinical decisions.

1 Background

Real-time Health Care Decision Support. Access to real-time and meaningful data at the point of care is one of the biggest challenges of the 21st century. In healthcare, a real-time clinical decision support (CDS) system that is shareable across healthcare delivery settings over large geographic regions plays an important role in improving outcomes. CDS technology has been demonstrated to improve the quality and safety of patient care, and is believed to...
be an integral component in improving outcomes [6, 7]. With a CDS system, caregivers can make treatment decisions for a patient based on real-time data. A real-time CDS system brings together data from multiple sources and provides tools for clinicians to access and analyse such data. Properly implemented, it enables patients and caregivers to make timely and informed clinical decisions.

SNF Study is a multi-center clinical trial that focuses on the immediate post-acute patient with HF (i.e. 30 days post hospital discharge) and the immediate post-SNF discharge (i.e. 60 days post SNF admission). The study focus on a vulnerable time period when the patient transitions from one care setting to the next. In addition, patients with HF can have a change in condition fairly rapidly. Therefore, a real-time CDS can be highly beneficial.

**DCDS: Data Capture and Decision Support System.** DCDS is built on the OnWARD [8] framework, a dynamic, secure, rapidly-deployed, and web-based form generator supporting data capture for large-scale multi-center clinical studies. For example, it has been successfully deployed in other clinical trials such as the Heart Biomarker Evaluation in Apnea Treatment trial (HeartBEAT,1RC2HL101417), a multi-institution Phase II clinical trial funded through the American Recovery and Revitalization Act (ARRA). OnWARD is developed using the agile methodology, involving regular communication with researchers/clinicians to collect requirements and update/improve the system progressively. OnWARD automatically translates structured specification of data format into web-based input forms without requiring deeper technical expertise. Thus, OnWARD can be quickly deployed and customized for different types of studies. DCDS inherits the following features from OnWARD:

- Flexible backend database selection. DCDS has an independent backend relational database designated by an investigator to support data entry, retrieval, and validation. This feature allows easy deployment/reuse of existing databases as well as providing complete control by the investigator.

- Dynamic form generation. For every data form, the data types, ranges and data distributions (such as text box, drop-down list) for all data entry points are transcribed into an XML file with specific format, which is in turn translated into an electronic data entry form by DCDS.

- Input validation. To ensure data quality, mechanism for validations on data types and ranges are implemented in DCDS at the time of data entry. Uniqueness of patient identifier are validated to avoid duplicated IDs among different patients.

- Branching logic. This is used to skip certain questions depending on the answers to previous questions. For example, in the medical history form, if the user chooses “no” for the question “Myocardial Infarction,” the next question “If yes, date of most recent MI” will be hidden from the user. The branching logic skip patterns are specified as metadata in the format of (question; condition; value; skip-question; default-value).

**Trial of Heart Failure Disease Management in Skilled Nursing Facilities.** HF challenges our clinical management skills because of unpredictable exacerbations, frequent hospital utilization, and complicated therapeutic regimens [9]. HF patients discharged to SNF have a higher rehospitalization rate and mortality than those discharged home [10], which is surprising since SNF are a major site of transitional care, from hospital to home, for older adults. The SNF Study is a randomized trial in which SNF physicians are randomized to either the HF-DMP managed by HFNA or usual care (UC). Quality initiatives to improve HF care in SNF can be vital to reduce rehospitalization and improve health outcomes for older adults. Although HF-DMP have been shown to reduce rehospitalization in other settings [11, 12], the effectiveness of HF-DMPs has not been studied in SNF. Since the SNF environment is highly regulated and programs are increasingly influenced by financial aspects, HF-DMP must both benefit patient outcomes and be affordable.

Research in SNFs is uniquely challenging. Each facility is its own entity with its own policies and procedures. Although multiple participating SNFs may be owned by the same company, there is considerable variability. A web-based system is exceedingly important to ease data collection burden at multiple sites with flexible software and computer access for study staff.

2 Challenges of DCDS

The HF-DMP is based on best practices for HF care, including documentation of left ventricular ejection fraction (in the form medical history), symptom and activity assessment, weights and dietary surveillance three times a week, medication titration, patient and caregiver education, discharge instructions, and a 7 day follow up visit post-SNF discharge. A specialist trained in Heart Failure Nurse Advocate (HFNA) works closely with physicians and ensures
fidelity of the intervention. Researchers in the SNF Study designed a wide variety of data collection forms to capture patient data in both UC and the HF-DMP group. Those forms fall into three categories (Figure 1): red color for shared forms, green for HF-DMP, and yellow for UC.

- Forms which are used once for baseline descriptive data, such as patient demographics and medical history.
- Forms which are used to collect outcomes data, which clinicians are investigating change over time. Such form include the Kansas City Cardiomyopathy Questionnaire (KCCQ) and Self Care in Heart Failure Index (SCHFI) forms. Clinicians collect the data at baseline and then at the end of the study after 60 days. The changes in scores for health status (KCCQ) and a patient’s ability to manage their heart failure (SCHFI) will help illustrate the usefulness of HF-DMP.
- Forms of HF-DMP, which are used 3 times a week and are interactive for assisting the HFNA in making clinical decisions by using the research protocol. These forms help the HFNA track HF signs, symptoms and weight so as to guide medication management. This kind of data informs the Research Nurse (RN) how to improve a patient’s treatment and intervene if the patient is decompensating from HF. The RN can make treatment recommendations if indicated by the protocol. The UC group has the same data collection forms, but the data is collected by chart review and there is no interactive component in UC. The HFNA is not involved in UC data collection.

Although the general OnWARD framework can fit most of what the SNF Study needs, there still remains a set of challenges. A fundamental challenge in the field of clinical decision support is to determine what characteristics of data make them effective in supporting particular clinical decisions. In HF care, patient’s weight, clinical symptoms, and medication titration are significant in evaluating the status of patients and determine their response to medical therapy. Whether the system facilitates appropriate access to such data affects the effectiveness in supporting clinical decision making.

2.1 Trending Graph Generation

Tracking patient’s weight, signs and symptoms is important for determining the clinical status of patients and gauging their response to medical therapy. The HFNA protocol weights on each patient 3 times a week and recommends
a dietary strategy based on the available SNF menu. The HFNA also assesses patient’s symptoms and activities to assist treatment decisions. However, isolated symptom or weight measure such as “swelling,” “dyspnea,” and “70lbs” is useless when it is evaluated beyond the context of a patient’s condition over time. In this case, trending graph generation is needed for DCDS. Using clinical data collected from patients at bedside, graphs are automatically generated to allow the HFNA to compare changes over time. Trending graphs provide visual clues about trends, gaps and clusters and compare multiple data sets at once that words and equations cannot offer. Based on trending graphs, study nurse can quickly draw conclusions and follow protocols to intervene on a patient’s change in condition in a timely manner.

2.2 Auto-calculation

In SNF Study, several forms require the calculation of scores based on patient surveys such as Demographics, KCCQ, Frailty, and Charlson Commodity Index. These scores form the basis for understanding patient conditions and information about the dynamics of disease when collected over time. The calculation of clinical scores has previously relied on Microsoft Excel, which calculates the scores based on data entered by study staff. Calculations in Excel needs cumbersome operations, which made errors easily, such as referring to the wrong cells or choosing the wrong data area. Automating the calculating process, which is a vital component of DCDS, not only saves time, but also avoids errors.

2.3 Medication History

Medications play a pivotal role in heart failure management. National guidelines for HF-disease management clearly delineate medications and target doses that will result in the best patient outcome [13]. Titration of medication can indicate improved quality of care for patients. Titration of diuretics in response to volume overload can help avoid HF decompensation and hospital admission. Medication titration according to pre-determined protocols is an important part of the intervention. Therefore, it makes sense to record medication changes occurring in response to the HF-DMP and to UC, and subsequently determine the effect medication titration has on outcomes. A medication history database for tracking, analyzing, and monitoring is another key component of DCDS.

The weekly medication form in SNF keeps daily records of medication name, dosage and frequency of every medication for each patient. The unique features required of this form are as follows:

- Auto completion for medication names. Due to the fact that medication names are often complex terms, clinicians can make errors when inputting medication information. The addition of real-time auto completion on medication names is a necessary functionality to improve efficiency and reduce error in the weekly medication form.

- Auto retrieval of medication history. Medication regimens can be lengthy and capturing medication information for each visit can be cumbersome. Since many medications prescribed to a patient infrequently change substantially, it is efficient for clinicians to directly modify the current medication information using the medication information of the previous day.

- Easy modification of medication. Based on medication information from the previous day, clinicians may add or delete medications to the current medication list. They may also need to modify the dosage or frequency of an existing medication. A user-friendly data entry interface which facilitates history-based medication information capture is required.

- Tabular form. In contrast to list form, a tabular medication form which allows structured collection and representation of information about medication, dosage, and frequency through a dynamic interface is desirable. Moreover, since the total number of medications is not predetermined, DCDS needs to be able to dynamically add or delete rows from the table.

3 System Design and Implementation

In this section, we present the design and implementation of DCDS. Initial development of the DCDS workflow was accomplished with close collaboration between the SNF investigator team and the DCDS development team. Figure 2 illustrates the workflow, including data collecting, decision support, and care delivery. In the SNF Study, patient’s clinical information is collected at bedside by the HFNA, and processed through DCDS to generate data reports to support HFNA management strategy.
Figure 2: Overview of our Real-time Data Capture and Personalized Decision Support System. Nurses collect patient information and conduct a clinic assessment. Data is ingested into DCDS. The system then generates timely clinical reports of the patient, such as weights, symptoms, and medications history. HFNA uses these reports to support clinical decision making. The clinician then delivers care by making changes to patient’s HF management if necessary. The HFNA captures changes to clinical care in DCDS. This work cycle is repeated during a patient’s entire SNF stay.

DCDS is developed using the Ruby on Rails web application development framework. It involves three seamlessly integrated modules: Data Specification, Data Capture, and Decision Support. Data Specification is the key step before Data Capture. All the metadata forms are specified to facilitate dynamic form generation. The metadata specification is captured in Excel sheets, which are converted to XML specification files by DCDS. The XML files then are directly loaded to the relational database of DCDS. After loading of the data specification, DCDS dynamic form generation engine automatically displays the form content for Data Capture. In the review and reporting phase, DCDS’ Decision Support component generates customizable reports using real-time data in the database.

3.1 Data Specification

Data Specification involves three types of metadata: variables, value domains, and permissible values. A form consists of a collection of variables, whose specification includes datatype, distribution, descriptive label, and value domain. The value domain of a variable has associated permissible values that determine which values are allowable for the specified domain. For a categorical value domain, each of its permissible values serves as an option of the drop-down menu list; for a numerical value domain, its permissible value specifies the valid range.

3.2 Data Capture

DCDS has built-in data capture features, such as data validation, multiple data type supporting, and branching logic. For the unique needs of SNF Study, we provide the following functions to improve the efficiency of data capture.

**Auto Calculation for Clinical Scores.** Calculation formulas are stored in the backend database. Each formula involves a number of questions in the same data entry form, divided into two groups: conditional questions and result questions. As a user enters data in the conditional questions from the web interface, a javascript function is triggered and automatically fills in the result entry with the correct value based on the formula. For example, in the KCCQ form (Figure 3), 10 scores are filled automatically based on the answers to the previous questions. This avoids the additional needs for using separate Microsoft Excel or any other calculation tools.

**Auto Completion of Medication Names.** A backend medication dictionary is imported beforehand. We extract the generic medication names from RxNorm [14], a normalized naming system for generic and branded medications provided by the National Library of Medicine (NLM). We built the “auto completion” function this way: When a user enters the first letters of a medication name, the system automatically pops up a list of all partially matched medication names.
names containing the input string as a substring for the user to choose.

![KCCQ Auto-calculation](image1)

**Figure 3: KCCQ Auto-calculation.**

![Medication Form](image2)

**Figure 4: Medication Form.**

**Auto Retrieval of Past Medications.** In the SNF Study, medication information is recorded daily. The automatic retrieval process is divided developed in these steps: When a user creates a new medication record, DCDS finds the previous record by date, and then automatically retrieves the medication history from that record and displays it in a table, as shown in Figure 4. Pre-processing is needed to make this process efficient: if a medication was labeled “discontinued,” it would not be displayed.

**Easing the Burden on Medication Modification.** After retrieving the medical information from the previous record, DCDS provides users with the ability to make changes to each medication to generate a new medication order for the current day. They can increase or decrease the dosage or frequency by choosing the appropriate option in the drop-down menu. There is also an option named “Discontinue” in the drop-down, which is used to remove the corresponding medication from the order. The process of adding a new medication goes as follows. First enter in the medication name with the help of auto-completion; then click the “Add” button; then a new row with the medication name is added to the table. Other fields capturing dosage or frequency information need to be filled by the user. DCDS preserves each medication list from each day, which means that even if a medication is discontinued one day, the medication history is still preserved for the previous day. In this scenario, it does not overwrite the medication list and HFNA truly follows medication management trends.
3.3 Decision Support

To support clinical decision making, we designed several reporting tools to help data tracking and analyzing. Tabular form of medication history along with the changes in weight and symptoms provides an intuitive grasp of patient’s condition and helps HFNA deliver appropriate care.

**Tabular View for Medications.** The form generation engine in DCDS provides two basic layouts: List View and Tabular View. The Tabular View displays a multi-row and multi-column layout. The medication form in SNF has eight columns: date of visit, medication, change, indication, dosage, unit, frequency, and change based on HFNA recommendation. Among the eight fields, “Date of Change” and “Medication” are two key fields. The “Change” field contains five options: no change, change in dosage, change in frequency, add new medication and discontinue. “Indication”, “Dosage”, and “Unit” fields are text fields. The “Frequency” field records medication frequency, which includes four options: once a day, twice a day, three times a day, or as needed. The last field “Other Frequency” is a text field as a supplement to the standard “Frequency” field.

![Medication History](image)

Figure 5: Medication History.

**Weight-Trending Graph.** Tracking of weight over time is a key aspect of SNF Study. Doctors track weight closely for patients with HF because weight gain due to fluid retention is an early sign of decompensated HF. The information is used to prescribe medication to patient and also to review response to medications. For the purpose of usability and analysis, the graph needs to be as interactive as possible. We use Google Charts along with Javascript to fulfill this purpose. Google Charts fit our requirements and have proved useful. Javascript is used to dynamically show and hide the graph.

![Weekly Weight](image)

Figure 6: Weekly Weight.

Weight-Trending graph has the following elements:

- Intelligent Y-axis. The Y-axis of the graph, which represents the weight over different time intervals, adjusts
itself according to the present value range of weight so that small changes in weight are easily visible.

- **Time Interval Adjustment.** Graphing range can be switched from 1 day to 1 year according to the length of the period during which a patient is under observation. This way the HFNA can zoom in and out at different time intervals.

- **Mouse Pointer Value.** At any point, the graph shows the value of the weight with the date corresponding to the current position of mouse pointer.

**Symptom Analysis.** Various symptoms can be present for different patients in different ways. To know how these symptoms are changing over time for a particular patient proves to be of great importance in SNF Study. It makes sense to implement a chart or table that gives all the required information about various symptoms of a patient in a concise manner. This not only saves HFNA’s time but also creates simplicity and improves recognition of changes in a patient condition.

![Patient Symptom Report](image)

**Figure 7: Symptoms.**

- **Tabular Form.** About 12 different symptoms are analyzed. Displaying such information in a chart may create confusion. Therefore it makes sense to follow a tabular form approach. It simplified the design and served the purpose beautifully.

- **Symptom Progress.** ‘+’ indicates that the symptom has increased, ‘-’ indicates that the symptom has deceased, ‘=’ indicates that the symptom is the same as the previous, ‘N/A’ indicates that symptom progress is not available, ‘X’ indicates that symptom was not previously recorded so there is a lack of information to be shown.

- **Colors.** To make the table even more interactive, in addition to different symbols, we added appropriate background colors. Green indicates an improvement, red a worsening, yellow the same, gray no records, and white not available.
4 Results and Evaluation

In total, we created 22 clinical forms containing 584 questions and 175 branching logic in DCDS. The basic demographic form assigns a study identifier for easy tracking, which is used as the universal key for other 21 forms. We stored 4103 medication names in our backend database to enable auto completion of medication entry and 5 calculation formulas: Demographics, KCCQ, Frailty, SCHFI and Charlson Commodity Index. DCDS is presently used in 43 facilities in the Denver area. The SNF Study is still enrolling and currently (March 2016) it has enrolled 365 patients including 44 in Cleveland.

<table>
<thead>
<tr>
<th>Elements</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilities</td>
<td>43</td>
</tr>
<tr>
<td>Patients</td>
<td>365</td>
</tr>
<tr>
<td>Forms</td>
<td>22</td>
</tr>
<tr>
<td>Questions</td>
<td>584</td>
</tr>
<tr>
<td>Branching Logic</td>
<td>175</td>
</tr>
<tr>
<td>Calculation Formulas</td>
<td>5</td>
</tr>
<tr>
<td>Medication</td>
<td>4103</td>
</tr>
</tbody>
</table>

Table 1: Summary of Results.

4.1 Evaluation

Compared to traditional paper-based method of data capture, structured electronic data capture has several advantages. It reduces the data collecting burden, supports quality reporting, improves the capability of informing researchers, and ultimately, enhances patient care. An important benefit of DCDS is that through the VPN, users can access the system anywhere. SNF staff usually goes to many facilities, including patients’ homes. With easy access to the system, they can get on DCDS anywhere and do not need to return to the office to enter data.

In this section, we compare DCDS with Paper-Based method to evaluate the unique design of DCDS for HF-DMP study. We designed a survey to evaluate the two methods (paper froms vs DCDS), including aspects of entering, tracking, usability, customizability, consistency, and representation. We invited 3 professional research assistants and 2 nurses in the SNF Study to take the survey and collected their comments and notes about the two methods. Each participant did data entry using paper-based HF-DMP workflow (data was entered into DCDS later), then performed the same HF-DMP workflow on a different patient and directly captured data into DCDS.

In the figures: a) Entering denotes how long and easy to complete the forms; b1) Tracking(in medicatin and weekly visit) denotes how difficult to track the changes of symptom and weight; b2) Scoring(in KCCQ) denotes how difficult to track the changes of symptom and weight; c) Usability denotes overall usability; d) Customizability denotes the overall customizability; e) Consistency denotes how difficult to make sure information you write is correct; f) Representation denotes how well the information represented.

Figures 8, 9, 10 show the average ratings of three selected forms: 1) Medication, 2) Weekly Visit and 3) KCCQ. Each form was rated on 6 aspects and each aspect was measured on a scale from 1-10 with 1 being the hardest and 10 the easiest. Figure 8 shows that the ratings of DCDS are way ahead of the paper-based method in all aspects. With respect to the testers’ comments, “Didn’t have to retype out names for DCDS” and “Takes a long time with the amount of meds” for paper-based method confirmed that DCDS is well-designed, as a result of auto completion of medication names and tabular view of the medication history. However, in Figures 9 and 10, the results are different. Our DCDS is only leading in tracking and scoring, which shows a gap from our expectation. From the comments and notes for these two forms, we found several reasons. These include: “Papers are easy to carry and write down any notes,” “Booting up computers, login and access to web-based tools is cumbersome,” “Patients are less willing to expand on answers when sitting in front of computers, which seems to be a more formal interview.”

A detailed analysis revealed that participants preferred using DCDS in situations of cumbersome repeating input, long-
term tracking, or scoring. The results demonstrated that our design is practical and user-friendly. Another interesting finding is that patients are more willing to answer questions using a paper-based method. They prefer casually talking with a nurse rather than a formal interview in front of a computer.

5 Discussion

As we can see in the evaluation section, evaluators preferred using paper to record clinical information than using DCDS. It took more time to boot up the computer, connect to the VPN and sign into DCDS than to fill out a form. However, the results showed increased satisfaction with the ease of data entry, graphing for symptom tracking and auto calculation for clinical scores. The data processing done by DCDS made clinical decision making easier as data was presented in charts and graphic forms, making day to day comparisons of clinical data straightforward. In addition, the advantages of DCDS are not obvious from one or two patient visits. As researchers become more familiar with DCDS, their efficiency likely increases. In the short-term, especially for a study such as SNF, users may spend more time in using DCDS. However, the long-term benefit in ease of exploring and managing clinical information of patients is important. Since all data are captured in DCDS on each patient, it provides a rich dataset for real-time data analysis.

Health care data standards are another important aspect which we need to consider and integrate into our DCDS in future work. Using the guideline of data standards, such as HL7, we can use standard terminologies, data formats, methods and protocols for data exchange, integration, representation, and decision support.

6 Conclusion

In this paper, we present DCDS, a real-time data capturing and personalized decision support system, for SNF Study to investigate the effectiveness of HF-DMP in improving outcomes for older frail heart failure patients undergoing post-acute rehabilitation. The results show that DCDS is clearly superior to existing manual approaches in data retrieval, data visualization, and easing the burden for cumbersome data entry.

Acknowledgement

This publication was made possible by NHLBI award R01HL113387. This work is also supported in part by University of Kentucky Center for Clinical and Translational Science (Clinical and Translational Science Award UL1TR000117).

References

Improving Veterans Access to Care through Telehealth and Local Fire Departments

Hamed Abbaszadegan, MD, MBA1,2, Umar Iqbal, MD2, J Edward Maddela, MD2, Khaleel Hussaini, PhD2, Manolo Moneda, RN, BSN, MBA1, Shakaib Rehman, MD1,2

1Phoenix VA Healthcare System, Phoenix, Arizona; 2University of Arizona College of Medicine, Phoenix, Arizona

Abstract
The VA Healthcare System has long been on the front lines of cutting edge telehealth care delivery. Fire departments have also been changing their approaches to emergency care by including new health and medical services. Pairing these two existing services will provide Veterans instant access to healthcare providers. The Clinical Informatics Department at the Phoenix VA recognizes this and has collaborated with local fire departments to reach out to Veterans in need. A pilot study with the City of Chandler Fire Health & Medical (CFHM) was initiated to improve healthcare access for Veterans residing in Chandler, Arizona. CFHM already has infrastructure to assist the VA through its Community Paramedic (CP) Program. Two algorithms of care were developed for the CP program to provide instant access for their patients identified as Veterans through an on-site, real-time telehealth visit with a VA Nurse Practitioner (NP). Additionally, these NPs will deploy community paramedics to Veterans with a high Care Assessment Need score within the service locale of CFHM in an effort to prevent ER visits and readmissions. These algorithms, the telehealth technology, and 6 months of data will be presented for this system's demonstration.

Problem
Access to care is a national problem in all healthcare systems due to provider shortages. Improper use of emergency room resources for non-emergent needs has become a national health care crisis. Often, lack of care coordination has been a major limitation. The Phoenix VA Healthcare System has been working diligently at addressing this problem over the last few years. Telehealth and clinical informatics are a major solution to leverage alternative means of enhancing care delivery.

Purpose
The purpose of this pilot is to identify Veterans in the City of Chandler who may need nonemergency services, and reach out to them in a predictive fashion prior to emergency scenarios. We aim to demonstrate that combining the resources of two service industries, a local VA health system and a city fire department through telehealth can potentially positively impact access to care, care coordination, and prevent care crisis scenarios.

Description
In our innovative program, there are two models of which CHFM and the Phoenix VA Health System provides telehealth remote care to Veterans. In the first model, when a Veteran activates the 911 system, CHFM responds in the usual fashion for an emergency call. If the call is deemed non-emergent, the Veteran will be cared for by CFHM through use of telehealth interface to connect Veterans to the VA instantly. Veterans can be seen remotely by a VA employed NP who has full access to the EHR with abilities to order medications, request consults, and even request an earlier in-person appointments. If the call is emergent, the Veteran is transported to an ER per prior protocol. The other model focuses on non-emergent follow up care via telehealth. Using the Care Assessment Need (CAN) score, a predictive analytic tool predicting high risk/high utilizers, patients are enrolled into this pilot project program. The CAN score is based on factors reported including but not limited to: prior ED utilization rates, hospital admission and readmissions in the last 90 days, age, and a number of chronic disease medical problems. This patient data is generated daily based on the zip codes where CHFM provides service. Once the data is generated, demographic information is shared with CHFM to coordinate care with the VA’s NPs scheduling Veterans for a telehealth visit.

Degree of Service Deployment
The prototype has been running for 3 months at the time of proposal submission. We are successfully seeing approximately 5-7 Veterans per week in conjunction with Chandler Fire Department and collecting data. We plan to present the 6 months’ intervention data at the meeting. Preliminary data is showing promising results, and all parties involved including Veterans, Chandler Fire, and the VA are providing positive feedback to this leading edge system that promotes an innovative method of healthcare delivery.
Carelign©: Upgrading from Outdated Paper Handoffs to Real-Time, Interdisciplinary Handoffs
Subha AiranJavia, MD1,2, Jacqueline Soegaard, SB1,2, Jillian Olsen, BA2, Geoffrey Bass, MD, MBA1, Damien Leri1, Glenn Fala1, Rich Urbani1, William C Hanson, III MD1,2
1University of Pennsylvania Health System, Philadelphia, PA; 2Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

Abstract The acute care of patients involves a multitude of moving parts – from teams of physicians, nurses, and interdisciplinary providers, to an array of tests, medications, and clinical information systems. Physician teams often use paper “handoffs” not only for transitions in care, but also to organize these parts to work together. These documents, however, are outdated soon after being printed. Moreover, valuable information written on these printouts seldom gets communicated promptly to other providers, if at all. To improve the quality of transitions, interdisciplinary care, and clinician efficiency, Penn Medicine created Carelign ©: a responsive web application with real-time clinical data and an interactive, editable, electronic handoff with full mobile functionality. Designed via close collaboration between physicians and developers, Carelign © gathers information from over 10 different systems into one clinically intuitive display. Our real-time, shared, and easily accessible handoff interface provides a timely view of the patient’s problems and the team’s plan for each. Viewable by all providers, including not just clinicians, but also nurses, technicians, phlebotomists, social workers, clinical nutritionists, and others, Carelign© is transforming clinicians’ daily workflow and interdisciplinary team interaction.

Description Carelign © is a homegrown, responsive web application designed by frontline clinicians, residents, medical students, interdisciplinary providers, and developers. Device and operating system agnostic, Carelign © is accessible from anywhere hospital computers, workstations on wheels, smartphones, tablets, or personal laptops as long as the device is connected to the hospital’s secure network via WiFi or VPN. Carelign © was developed and launched in 2 phases: Version 1 (V1) in August 2014, provided real-time access to objective data, and Version 2 (V2), implemented January through July 2016, adds an iterative, collaborative plan of care used in handoffs and to create progress note assessment and plans. Utilizing HL7 messages, direct database queries, and web services/APIs, Carelign © pulls information from our many different information systems: ADT (demographics and visit information), Cerner (lab data), Radiant (imaging), Navicare (patient location), Epic (provider and encounter data), Sunrise Clinical Manager (care team, vital signs, allergies, and medications with administration record), and others. Accessing all this critically important data previously required going to multiple different programs. Now it is available at any time, through an intuitive display optimized for smartphones. Initial time motion observations revealed that accessing Carelign © on mobile devices was 22 times faster than logging into the inpatient EHR on a workstation on wheels. Although originally designed for inpatient use, Carelign © has become widely used by our home care clinicians and interest in other “off-label” uses continues to rise: from food services using it to verify diet orders at meal delivery to workgroups aiming to provide a better foundation for interdisciplinary discharge planning.

Beyond data review, V2 adds handoff functionality that is arguably more transformative, moving away from free text boxes to discrete, collapsible data elements organized in a way that facilitate rounding, critical thinking, and list management. Key features include the ability to archive information, tag items as anticipatory guidance, and view qualitative content arranged by problem or by task. Carelign © also provides extensive functionality for list-based management of patient and team tasks. We have implemented Carelign © V2 across almost all inpatient clinical services at the Hospital of the University of Pennsylvania, Penn Presbyterian Medical Center, Pennsylvania Hospital, and Penn Medicine at Rittenhouse, with the goal of transitioning the last few services by the end of July 2016. Almost two years after the launch of V1 and 6 months into V2, over 3,000 unique patient records are viewed weekly between >23,400 sessions amongst >3100 unique users: 40% physicians & advanced practitioners (including 7% attendings), 45% nurses and 13% other clinical users (pharmacists, therapists etc). Provider surveys comparing analogous medicine teams using Carelign © compared to our previous handoff system, demonstrate that residents report significantly higher satisfaction and safer written handoff content with Carelign ©, as well as a near-significant improvement in workflow efficiency.

By upgrading our handoffs and care coordination from a paper-based system to Carelign ©, a shared and interactive application with a fully functional mobile interface, we are improving user satisfaction and efficiency, subjectively improving written handoff safety, and hope to demonstrate an improvement in interdisciplinary care.
Assessment Center API: A Software Component Model for the Integration of Patient Reported Outcomes (PRO) into Clinical Care

Michael Bass, MS¹; Paul Harris, PhD²; Rob Taylor²; Justin Starren, MD, PhD³; Joshua Spuhl, MS⁴; Jimmy Johnson⁵; Jason Guattery, MS⁵

¹Department of Medical Social Science, Northwestern University, Chicago, IL ²Department of Biomedical Informatics, Vanderbilt University, Nashville, TN ³Northwestern University Clinical and Translational Science Institute, Chicago, IL ⁴University of Utah Health Care, Salt Lake City, UT ⁵Washington University School of Medicine, Saint Louis, MO

Abstract. The Assessment Center API https://www.assessmentcenter.net/ac_api¹ is an easy to implement system that enables system integrators to offer Patient Reported Outcomes Measurement Information System (PROMIS) ² instruments within their patient-facing portals and applications. In this system demonstration, we will first provide an introduction to Item Response Theory (IRT) and computer adaptive testing (CAT), the underlying principal in the administration and scoring of PROMIS instruments. Next we will highlight the key features and workflow of this API system focusing on the integration with the API endpoints. Results from four case studies that have implemented the API will be presented. This will demonstrate the diverse range of approaches used in collecting PRO data in both clinical and research-based. The four test cases will come from the following systems: REDCap³ developed at Vanderbilt University, mNPOR developed at Northwestern University, mEVAL developed University of Utah and WUPRO developed at Washington University School of Medicine.

Background and System Description. Item response theory (IRT) and computer adaptive testing (CAT)⁴⁵ have been used successfully for decades in education, licensure, achievement testing, personality assessment, and selection of military personnel. By only administering questions targeted to the subject’s ability or trait level, testing times can be cut in half while at the same time improving overall test reliability⁵. Most recently IRT and CAT have been applied to the domain of health-related quality of life in the form of patient reported outcomes (PRO)⁷ which provide information about a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a physician. The reduction of patient burden is the primary motivation for implementing an IRT-based administration and scoring algorithm in the PROMIS instrument framework. These algorithms were initially developed and deployed in Assessment Center⁸, which provides a turnkey solution for researchers to conduct PRO research. This platform is not suitable for clinical research and workflow because Assessment Center is based on a “software as service” (SAS) delivery model and would not satisfy security or workflow requirements of most clinical settings since patient data would be stored outside of an organization’s infrastructure. For these reasons a portable component model was developed for the use case of administering PROMIS instruments which can be integrated into an existing clinical system. The API is based on a RESTful architecture that follows a lab or order-based clinical workflow/paradigm. First order(s) for a PRO assessment are placed to the API. Next, the PRO assessment is administered to the patient, representing the processing of the order(s), followed by the retrieval and integration of results from the API. The API was initially released in 2012 and has been successfully integrated into both research and clinical based systems that have patient-facing interfaces. We will present results and key integration points from a selection of production clinical and research systems that have implemented the API. We will also present the usage of the API endpoints, by demonstrating the calling syntax and returning formats. Finally, an example Javascript application will be presented that demonstrates how a clinical patient facing workflow, intended to capture PRO data, can be built using this API system. The most recent work on this project has been the porting of the API system to architecture suitable for embedding in mobile platforms. Demo applications in both Android and iOS will also be presented, this will highlight the versatility of the API.

References
Integrating a Dialogue System into a Virtual Patient Consultation

Leonardo Campillos-Llanos, PhD, Dhouha Bouamor, PhD, Éric Bilinski, Anne-Laure Ligozat, PhD, Sophie Rosset, PhD, Pierre Zweigenbaum, PhD
LIMSI, CNRS, Université Paris-Saclay, Orsay, France

Abstract

We present a prototype of a dialogue system simulating a virtual patient in a medical consultation. The system is integrated in a tool aiming at providing continuing education to medical doctors, and the current version supports the French language. The conversational agent developed by LIMSI allows medical students and practitioners to train and improve their social and attentional abilities during the history-taking stage of a consultation. We will explain the difficulties found in medical dialogue processing. A first challenge is to process specialized terms and match them with the patient’s clinical record defined by the clinical case designer: natural language understanding of doctor’s utterances. A second challenge is to generate utterances which simulate patient speech: taking a layman’s perspective. Additionally, the system needs to cope with case-specific questions and to handle dialogues for new clinical cases in a robust way. We will outline the strategies and resources for managing linguistic and terminological variation to carry out these tasks. We will show the methods we use to achieve a natural user interaction, by addressing the structural variation of users’ interventions and taking into account dialogue phenomena such as anaphora and ellipsis.

Description and purpose

Virtual patients (VPs hereafter) allow trainees and practising doctors to develop expertise in diagnosis or clinical reasonings. Patient-Genesys, an interdisciplinary project managed by Interaction Healthcare, aims at developing a computer tool for creating digital clinical cases to provide beginning and continuing education to health professionals in French language. The system works as follows. On one side, a medical trainer creates the clinical case to be simulated by selecting a scenario and an avatar, and by providing a description of the VP’s health state expressed in plain text as a clinical record. On the other side, a trainee is placed in a situation where s/he has to perform a consultation with the VP. This is represented by an avatar with 3D animation in real time (Interaction Healthcare). The trainee can ask questions through text input to obtain information from the VP. LIMSI’s natural language dialogue system analyses the user’s input and replies according to the clinical record. The reply is output by a text-to-speech system (Voxygen) with lip synchronization. Lastly, the trainee may formulate a diagnosis or verify drug interactions (VIDAL), and receive feedback. The Angers University Hospital (CHU d’Angers) provides medical expertise to the project.

Conclusion

The system has been tested by project partners and at public demonstrations. The text dialogue interface which is the focus of this demonstration is available online for the 3 cases developed to date: anesthesia, hypertension and pneumopathy.

Acknowledgments

This work was partly funded by BPI through the FUI project grant PatientGenesys (F1310002-P).

References


http://www.audiosurf.net/pg5c/select_case.php
ExplodeLayout: Comprehending Patient Subgroups in Large Networks
Bryant Dang BS\textsuperscript{1}, Joseph Mathew MS\textsuperscript{1}, Tianlong Chen BS\textsuperscript{2}, Suresh K. Bhavnani PhD\textsuperscript{1}
\textsuperscript{1}Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX;
\textsuperscript{2}Department of Physics, University of Houston, Houston, TX

Abstract
Networks have been used to successfully identify and comprehend patient subgroups based on their characteristics (e.g., comorbidities or genes) with the goal of designing targeted interventions, a cornerstone of precision medicine. However, current network layout algorithms often fail to reveal patterns in large and dense networks despite having significant clustering. We therefore developed an algorithm called ExplodeLayout, which exploits the existence of clusters to automatically “explode” a traditional network layout with the goal of separating overlapping clusters to enhance their comprehensibility. We demonstrate the use of this algorithm to visualize a large dataset extracted from Medicare consisting of readmitted hip-fracture patients and their comorbidities, which enabled clinicians to comprehend patient subgroups and their comorbidities, and to infer mechanisms related to hospital readmission.

Introduction
Although network visualizations have enabled researchers to identify and comprehend patient subgroups based on their characteristics (e.g., genes or comorbidities),\textsuperscript{1} current layout algorithms such as Fruchterman–Reingold (FR) often fail to reveal such subgroups despite the networks having significant clustering. For example, Fig. 1A shows a bipartite network consisting of all 30-day readmitted hip fracture patients (n=6150) extracted from the 2010 Medicare database that had at least one of the 8 significant comorbidities shown. Unfortunately, the network layout generated by FR is difficult to comprehend due the overlapping clusters despite (1) the network having strong clusteredness (co-clustering modularity\textsuperscript{2}=0.440) that was significant (p<.001) compared to 1000 random permutations of the network, and (2) the use of distinct colors to denote each co-cluster of patients and comorbidities. Here we demonstrate ExplodeLayout an approach that exploits information about such clusters to enhance their comprehensibility.

Method and Results
ExplodeLayout is implemented in R, and requires as input (1) coordinates of network nodes generated by a traditional force-directed algorithm such as FR, (2) the number and members of co-clusters generated by a modularity algorithm\textsuperscript{1} (standard in most network applications), and (3) a user-selected circle radius using a scroll bar to determine the distance by which the node clusters are moved. The algorithm uses these inputs to (1) generate an imaginary circle for the given radius with n equidistant points, where n=number of the clusters, (2) move the nodes within each cluster by the same distance and angle such that the centroid of that cluster coincides with one of the points on the circle. As shown in Figure 1B, this transformation effectively “explodes” the cluster locations by preserving the distances among nodes within the clusters, and by not preserving the distances among nodes across the clusters. This trade-off between not preserving inter-cluster node distances and having greater separation between the clusters enables a researcher to comprehend the relative size and topology of each cluster. An optimal layout is generated iteratively by the researcher to achieve a desired balance between cluster separation and layout compactness.

Layouts generated from ExplodeLayout have enabled important insights\textsuperscript{1} into the risk factors for subgroups of readmitted hip-fracture patients, and therefore could be an important complement to force-directed algorithms such as FR. However, for networks with a large number of clusters, the user would need to select a large circle radius to enable cluster separation, resulting in unused space in the center of the layout. We therefore implemented the ability to specify the number of nested circles enabling the central region of the layout to be filled and therefore still maintaining separation with compactness. In future research we plan to automatically suggest to the user the appropriate number of nested circles, which can be modified and therefore enable the rapid generation of compact and comprehensible layouts consisting of hundreds of thousands of nodes, and with many patient subgroups.

References
Automated Evaluation and Training for Interprofessional Education using Virtual Patients and Providers

Debajyoti Datta, MS1, John Owen, EdD2, Valentina Brashers, MD2, Casey White, PhD3, Laura E. Barnes, PhD1

1Systems and Information Engineering, University of Virginia, Charlottesville, VA; 2School of Nursing, University of Virginia, Charlottesville, VA; 3School of Medicine, University of Virginia, Charlottesville, VA

Abstract
Adverse events resulting from medical errors cause significant morbidity and mortality and cost billions of dollars annually. Communication failure within healthcare teams is a leading cause of medical errors. Interprofessional Education (IPE) has been demonstrated to improve healthcare team communication and prepare students and clinicians to deliver safe high-quality team-based collaborative patient care. Health professions schools and care delivery organizations (especially Accountable Care Organizations) are integrating IPE into their training activities across the learning continuum to meet new accreditation and licensing standards; therefore, educators need reliable outcome measures to assess whether learners are gaining essential team care competencies. Few validated objective observational instruments that measure these competencies behaviors are available, and their effectiveness in achieving rigorous learner assessment is limited. Furthermore, most current IPE training is performed with standardized patients making it very expensive and cost-prohibitive to many organizations. To address the deficiency in training tools for IPE, we designed a system specifically for IPE training, a platform for Virtual Patient and Provider Interprofessional Teamwork Objective Structured Clinical Examinations (VPP-ITOSCE). Virtual human patients have been used in training medical students in interpersonal communication skills and demonstrate a significant correlation between medical students interviewing a standardized patient and a virtual patient.

Specific Problems/Purpose
To develop an integrated platform that utilizes virtual patients and providers as well as state-of-the-art natural language processing to assess teamwork competencies. Currently this is being used for a Rapid Response Scenario for a patient with Chronic Obstructive Pulmonary Disorder (COPD). The patient is refusing treatment and it is upon the nurse to convince the patient to have the medicine. A virtual doctor is also present in the scene and the nurse is evaluated on the complete teamwork competency during the clinical examination scenario. The system will be evaluated based on the accuracy of question classification by the NLP engine and the response by the virtual agent, the grading comparison between a human evaluator and the automated system and the overall ease of interaction between the nurse and the system.

Key Features and Innovation
The demonstration involves nursing students interacting with a realistic virtual avatar on screen and using text input ask scenario questions. Behind the scenes, the system uses state of the art natural language processing using convolutional neural networks along with realistic animations and sounds to create realistic responses. The student is graded automatically based on pre-specified criteria. The system is built using a modular framework to ease extension to other scenarios.

Deployment
Currently we have conducted preliminary testing with a group of nursing students and will pilot test the system in the near future.
**Tiresias: Knowledge Engineering and Large-Scale Machine Learning for Interpretable Drug-Drug Interaction Prediction**

Achille Fokoue, MS, Oktie Hassanzadeh, PhD, Mohammad Sadoghi, PhD, Ping Zhang, PhD
IBM T.J. Watson Research Center, Yorktown Heights, NY

**Abstract**
Accurate prediction of potential Drug-Drug Interactions (DDIs) is a major requirement both in drug development and in patient care. In this demonstration, we present Tiresias, a system that takes in heterogeneous sources of knowledge as input and provides as output an accurate and interpretable prediction of potential drug-drug interactions. We will show various components of the system that perform knowledge curation from public sources of drug data, large-scale similarity analysis across all drugs using several similarity measures, and feature selection and efficient learning using a logistic regression model to not only predict potential DDIs, but also explain the reason behind each prediction. We will show an evaluation of the system using knowledge derived from an old drug database from 2011, verifying the predictions and explanations with recently discovered interactions and their evidence. The system includes APIs and interfaces that expose the outcome of each of the components to users and developers and enable integration within existing electronic health record and patient data management systems.

**System Description and Areas of Innovation**
In this demonstration, we will take the audience through various steps of building a scalable predictive model that can accurately predict drug-drug interactions for newly developed or existing drugs. Figure 1 shows the architecture of our system. Apart from the end result of DDI predictions, each component comes with an API that will be used to show the system’s features during our demonstration. First, we will show visualizations of several subsets of the knowledge graph constructed from various public sources of drug data [1]. We will then go over the similarity analysis component, which computes several similarity measures including but not limited to a measure based on chemical-protein interactome (CPI) profiles, measures based on mechanisms of action and pathways, and a measure based on Anatomical Therapeutic Chemical (ATC) classification system. The output of this component is used as input to a large-scale machine learning process that utilizes similarity matrices to generate features and adopts a logistic regression model to perform the prediction. A highlight of the system is the ability to interpret the predictions using a ranking of the most important features and how each similarity measure and source of knowledge contributes to the calculated confidence in a predicted DDI. A Web interface is designed using the APIs that allows the users to search for DDIs for a given drug and get explanations for each DDI that include details of similarity measures and contributing factors for the DDI.

![Figure 1. Tiresias architecture: knowledge curation, large-scale similarity analysis and learning, and powerful APIs.](image)

**Current Deployment:** Our system is fully implemented and the APIs are currently made available on an internal version of IBM Watson Developer Cloud [2] at Alpha maturity level. Our large-scale similarity analysis and learning is implemented on top of Apache Spark. We have performed thorough experiments that clearly show the effectiveness of our method in both predicting new interactions among existing drugs and among newly developed and existing drugs [3]. More details of our methodology, how it compares with state-of-the-art in DDI prediction, and detailed experimental results are presented in: [http://ibm.biz/adrrtechreport](http://ibm.biz/adrrtechreport).

**References**
DOT, an Interactive Reporting Dashboard for Visualization of Order Set Development and Utilization Patterns across an Enterprise

Nathan C. Hulse, PhD\textsuperscript{1,2}, Jaehoon Lee, PhD\textsuperscript{1,2}, Tim Borgeson, MBA\textsuperscript{3}

\textsuperscript{1}Intermountain Healthcare, Salt Lake City, UT; \textsuperscript{2}Department of Biomedical Informatics, University of Utah, Salt Lake City, UT; \textsuperscript{3}Cerner Corporation, Kansas City, MO;

Abstract
Successful computerized physician order entry (CPOE) implementations rely on building and maintaining a useful catalog of standardized order sets for use in a centralized knowledge repository. Collectively, these order sets represent a significant investment of effort and expertise in capturing and distributing best clinical practice throughout an enterprise. In order to address an important gap of understanding how order sets are both created and used in practice in our EHR installation, we have developed DOT (Data on Order set Transactions), an interactive dashboard aimed at informing content owners about how their order sets are being used in practice. The various views in the dashboard provide insight about many facets of knowledge utilization in the CPOE system, including evolution of the baseline templates, usage patterns by user group, location, and time, personalized order set library descriptors (including modification summaries), and line-item detail usage. The dashboard uses algorithms and pattern recognition to inform content authors about more relevant trends, including recommended optimizations to the order sets themselves (additions, removals, changes in selection status, etc). We anticipate that this type of utility will provide useful insight and feedback for those tasked with content governance and maintenance in CPOE systems.

Introduction
One of the important challenges of the information age in medicine is that knowledge evolves and expands quickly and keeping it current, useful and vibrant inside clinical knowledge bases requires significant investment. It is common for knowledge content to become outdated, and for content owners to become disconnected with knowledge consumers. Knowledge management lifecycle patterns characterized in information systems literature speak to the importance of taking the usage data that surround knowledge artifacts, compiling it into coherent units, and making that feedback available to knowledge workers through an active channel that facilitates learning. Given the significant investment in building out a new order set catalog in an ongoing EHR installation, we decided to build a visualization framework that gives content owners insight into how their order sets are being used.

System Description
In late 2013, Intermountain Healthcare and Cerner Corporation announced a strategic partnership, including an enterprise rollout of the Cerner EHR platform, and a joint development effort aimed at delivering novel informatics solutions for healthcare organizations. One of the strategic priorities of Intermountain’s Cerner installation focuses on implementing and measuring usage of best practice care throughout the system, with a particular emphasis on knowledge delivered in the form of order sets. In order to better inform content authors about how their content is used, we built DOT, an interactive dashboard that presents order sets’ uptake within the Cerner system. In this system, order set usage data is routinely extracted from the Cerner Millenium database and exported to Intermountain’s enterprise data warehouse. With this data, we implemented an interactive dashboard using Tableau software to allow content domain experts the ability to review how and where their content is being used. The dashboard provides several innovative views of this data including:
1. Visualization of order sets catalog ordering patterns against main order set templates, as well as aggregated ‘favorites’ data (personalized order sets derived from the original templates)
2. Summary views of individual and collective favorite order sets against the baseline template, highlighting common changes in content patterns (and potential opportunities for content refinement or user training)
3. Temporal view of content uptake over time, with dynamic filters for locations, providers, and user roles
4. Line item usage data, including detailed usage patterns of orders, implications of selection status, and modifications to order sentence detail
5. Actionable recommendations for content optimization, including additions/removals of orders, changes in selection status, order sentence detail

System Implementation
The system has been co-developed by members of Cerner and Intermountain’s informatics and analytics groups, with an early focus on content owned by the critical care, cardiovascular, and pediatric divisions. It is currently available for use in a beta release for specific clinical department analysts, with full enterprise release scheduled for fall 2016.

2117
Demonstration of GuideVue® mobile clinical guideline technology

Presenters:

- M Sriram Iyengar, PhD, Associate Professor of Biomedical Informatics, Texas A &M Health Science Center, Houston, TX
- Jose F Florez-Arango, MD, PhD, Research Assistant Professor of Biomedical Informatics, Texas A &M Health Science Center, College Station, TX
- Jack W Smith, MD, PhD, Professor of Biomedical Informatics, Texas A &M Health Science Center, Houston, TX

Abstract

Originally developed at NASA, award-winning GuideVue® is a complete technology for authoring, publishing, data reporting, and managing media-rich structured clinical guideline apps for healthcare providers and patients. The system is based on theoretical foundations of Felder-Silverman learning theory and Persuasive Technology.

There are five major components. The Author enables non-programmers to rapidly create media-rich mobile apps using an intuitive drag and drop, point and click interface where guidelines are presented as a series of steps in which each step can contain images, audio, text, video, data forms, communications, medical calculators and other capabilities. The Player apps are native IOS and Android apps into which guideVue apps are downloaded and executed. The Publisher provides one-click dissemination of guideVue apps for android and IOS platforms, the Reporter provides detailed usage reports, including compliance tracking, of guideVue apps with time/location stamping, and the Administrator enables system, app, and user management. Once downloaded into the mobile device, guideVue apps can function even without connectivity, making the technology very suitable for global health and developing country application, as well as supporting disadvantaged populations in the USA.

In this systems demonstration the presenters will provide a live demonstration of the entire system showing the creation of guideVue apps, publishing, executing the apps, compliance reports, and user/app administration. Existing apps for disaster response, global health, chronic disease, patent self-management will also be demonstrated. A summary of published research on Community Health workers and patients will be presented.

Specific Health Care Problem Addressed

Poor comprehension and adherence to clinical guidelines and procedures by health workers and patients has serious adverse effects on patient safety, risk of medical complications, length of hospital stay, healthcare costs, readmissions, and health outcomes. The problem is magnified among disadvantaged populations and in global health contexts where educational attainment and health literacy are low.

Deployments

Several pilot projects have been completed or in progress in Colombia, South India, Special Operations command, VA clinic (Pittsburgh), US Army Institute for Surgical Research, Shell Health, MD Anderson Cancer Center, The Ministry Health & Social Protection of Colombia, and others.
SMART on FHIR Platform for Interoperable Healthcare Applications

Joshua Mandel, MD1,2, Daniel Gottlieb, MPA2, Dan Fritsch, PhD3, Kenneth D. Mandl, MD, MPH2,1

1Harvard Medical School Department of Biomedical Informatics, Boston MA; 2Boston Children’s Hospital Computational Health Informatics Program, Boston MA; 3Polyglot System, Inc., Morrisville, NC

Abstract

We will demonstrate the core components of the SMART on FHIR Platform for Interoperable Healthcare Applications, providing an end-to-end view of the process of building and testing a clinical application using our developer tools and sandbox environment, and publishing the application in our publicly accessible SMART App Gallery where it becomes available for discovery. We will show how applications can be written against a standardized set of Meaningful Use data elements expressed using HL7 FHIR profiles that align with clinical terminologies including SNOMED CT, RxNorm, and LOINC, and then run with interactive patient selection and an OAuth-based access control mechanism. This approach enables secure integration of third-party, externally-hosted applications into diverse EHR and patient portal environments. We will discuss the platform’s adoption and deployment metrics as well as the challenges involved in app development. Finally, we will show how Meducation, a commercially available SMART app built against the platform, relies on these components to run unmodified within our own test environment as well as in commercial EHR systems.

Purpose

SMART Health IT is a standards-based technology platform designed to enable an ecosystem where third-party applications can integrate, unmodified, into diverse Health IT systems including EHRs, patient portals, and data warehouses. The platform is intended to address the needs of patient, provider and researcher end-users by stimulating a competitive landscape of apps that serve clinical care, wellness, and public health use cases. It is also intended to reduce the need for software developers to create custom data field mapping and security infrastructure each time an app is deployed in a new system. The system provides standards-based functionality for app development, testing, publication, and discovery, including a one-click “try it” button where potential users can run an app against a simulated EHR.

Deployment

The SMART platform has been deployed in multiple EHR vendor system including Cerner and Epic, and is in active clinical use at healthcare organizations including Boston Children’s Hospital, Duke Medicine and Intermountain Healthcare. We are also working with over a dozen EHR vendors through the Argonaut Implementation Program and the Sync for Science initiative to support wide-scale adoption.

Michael E. Matheny, MD, MS1,2, Dax Westerman, MS1,2, Laura Perlman, MS3, Josh Gieringer1,2, Xiaqian Jiang, PhD4, Claudiu C. Farcas4, Tara Knight5, Shuang Wang5, Amy Perkins5, Lucila Ohno-Machado, MD, PhD1, Bill Clarke5, Daniella Meeker, PhD5

1Tennessee Valley Healthcare System VA, Nashville, TN; 2Vanderbilt University Medical Center, Nashville, TN; 3University of Southern California, Los Angeles, CA; 4UC San Diego, La Jolla, CA 5Lincoln Peak Partners, Westborough, MA.

Abstract: The PCORNet program is a 4.5 year two-phase initiative that is supporting 33 health data networks using the PopMedNet™ data network software. PCORNet’s original version of PopMedNet™ was not designed to support peer-to-peer collaborations between networks, collective model estimation and specification, automated query execution approvals, or heterogeneous local software environments. We modified the PopMedNet™ portal and created a PopMedNet™ Adapter to enable these functionalities, which integrate with our evolving pScanner analysis software. This system demonstration highlights a collaboration between the PCORNet infrastructure development team (Lincoln Peak) and the pSCANNER CDRN analytics and software development teams to integrate a distributed analytics software suite with end-to-end functionality for use within the PCORNet infrastructure. Many of the system components are portable to other common data models, and the tools and modules will be discussed in this demonstration.

System Overview: We developed software modules that interface between the PCORNet data mart client adapter and the expanded pSCANNER analysis engine to address the need for (1) network-to-network collaborations, (2) menu-driven analysis specifications, (3) heterogeneous computing environments, (4) estimation of a network-wide model without case-level data transfer, and (5) visualization of analysis results. The resulting system includes the following features: regulatory approval and management user interface, horizontally partitioned model estimation with results identical to pooled data, data transformation and analytic execution plan programming representations based on widely used standards (Health Quality Measures Format [HQMF], Predictive Model Markup Language [PMML]) with translation modules to site-specific analysis engine, such as R or SAS, and results data visualization and graphing. The analytics core includes site-specific load-balancing (DAN), open source analytics and R-interface (OCEANS) analytics modules, and data visualization, table generation, and graphing components for study results.

Network to Network Collaborations and Protocol Approvals: We added a workflow to PopMedNet™ that enables any member of the network to propose a protocol to a subnetwork of collaborators through the portal. The data elements in a proposed protocol include specifications for the source data (e.g. PCORnet CDM, OMOP, custom data definition), specifications for analysis types (e.g. cohort discovery, privacy-preserving multivariate model estimation), requirements for local data and software environments, the type of local review process (manual vs. automated execution), and IRB and regulatory approval information. If a protocol is approved, multiple analytic queries can be submitted across the network under a single collaboration agreement.

Analytic Plan Representation & Translation: The software tools include a user interface that assists in developing data transformation rules from PCORNet CDM to flat analytic flat file by representing the query in HQMF, and including a translation modules at the site that translate the HQMF to SQL code, uses Hibernaner or SQLRender or other tool to translate between dialects of SQL, and passes the SQL into the site specific data processing engine. In the same way, predictive modeling analysis is developed within a UI, stored as PMML, and translated into site-specific analysis code through a translation engine, one for each flavor of common analysis platform (e.g. R or SAS). This allows data sites with different software environments to seamlessly collaborate on analysis protocols.

Privacy-Preserving Analyses: The DAN (Distributed Analytics Network) coupled with OCEANS (Observational Cohort Event Analysis and Notification System) provides an open source fault tolerant mechanism for processing analysis requests across available institutional resources. This includes a scalable mechanism for distributing requests across analysis engine resources, permitting asynchronous, parallel requests. OCEANS operates as both an analysis engine and workflow adapter, providing a strongly typed interface for analysis requests into other common engines, such as R. We also developed R and Java modules for distributed risk models (e.g., linear regression, logistic regression, Cox proportional hazards models, and support vector machine) that decompose optimization (using the Newton-Raphson Algorithm) to local private data to get faithful results as if models are trained on pooled data.

Data Visualization and Graphing: The software also includes Javascript based graphing functions developed within D3, a well-known graphing engine, for representation of prediction and association models, descriptive statistics, propensity score balancing and common ground assessment, and sequential surveillance methods (such as sequential probability ratio testing).

Deployment: The system is in use within the pSCANNER CDRN and each of the components is all open source and has been released with regular incremental updates. A portal link page to the open source releases and screenshots of the system is available at http://pscanner.ucsd.edu/Analytics/. PopMedNet™ was developed by the Department of Population Medicine at Harvard Pilgrim Health Care. Lincoln Peak Partners is the Software Development and Technical Partner for PopMedNet™.
Beyond SMART: Remote Decision Support with CDS Hooks

David P. McCallie Jr, MD, FACMI1, Joshua Mandel, MD2, Howard R. Strasberg, MD, MS, FACMI3, Scott P. Narus, PhD, MS4, Kevin Shekleton1, Peregrin Marshall5
1Cerner Corp., Kansas City, MO; 2Harvard Medical School, Boston, MA; 3Wolters Kluwer Health, San Diego, CA; 4Intermountain Healthcare, Salt Lake City, UT; 5RxREVU, Denver, CO

Abstract: FHIR®, HL7’s emerging data access standard, has rapidly gained attention as an appropriate model to standardize application programming interfaces (APIs) for access to HealthIT data. Concomitantly, the SMART® Health IT specification has become accepted as a way to embed “apps” into Health IT workflows. The combination of these two technologies, colloquially known as “SMART on FHIR”¹, has attracted interest as a way to create vendor-independent “substitutable apps” that can be plugged into a variety of EHR and other systems. SMART on FHIR currently provides only for synchronously-launched apps, where the user chooses to initiate the app via a link or a menu choice. The SMART group has recently proposed CDS Hooks, a set of specifications building on SMART and FHIR to provide a more deeply integrated user experience. The paradigm instruments a native EHR with an event model, triggering calls to remote CDS services when specific user activities occur (e.g., “prescribe a drug”, or “open a patient record”). These CDS services can respond with advice, alternative suggestions, and in-context app launch links that can be presented to the user in accordance with explicit user experience guidelines.

Purposes of the system: A growing number of organizations are now offering remote CDS services for deployment into EHR workflows. In the past, these remote services required custom code by EHR vendors, limiting the number of services that could be integrated, and indirectly suppressing CDS innovation. CDS Hooks provides a standards-based way to address these problems. We will demonstrate our pilot experience of the integration of multiple remote CDS services into a commercial EHR. The demo will include CDS integration triggered by multiple common workflow events, as well as a variety of modes of integration, including modal, non-modal, and ambient alerts. We will also discuss suggestions for enhancements that have emerged from the pilot testing process.

Deployment status: Pilot demonstrations of CDS Hooks have been shown in several public settings, and include support for multiple CDS service providers. Production deployment is expected in 2016.

References
CEDAR: Better Data Sharing Through the Authoring of Better Metadata

Mark A. Musen, M.D., Ph.D., Martin J. O’Connor, M.S., Marcos Martínez-Romero, Ph.D.,
Attila L. Egyedi, Debra Willrett, M.S., John Graybeal, and the CEDAR Team
Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

Abstract

The ability to find and access biomedical datasets that are stored in online repositories depends on the quality of the associated metadata. There is a growing set of community-developed standards for defining such metadata—often in the form of metadata templates. The practical difficulties of working with these templates are tremendous, however. The Center for Expanded Data Annotation and Retrieval (CEDAR) is developing technologies to assist in the management of biomedical metadata. By discovering patterns in existing metadata and by linking templates to biomedical ontologies, CEDAR is assisting the authoring of new, high-quality metadata. The availability of comprehensive and expressive metadata will facilitate data discovery, interoperability, and reuse.

System Description

CEDAR is building a suite of tools, known as the CEDAR workbench, that form a pipeline for authoring metadata. ¹

**Metadata Template Repository:** We have developed a standardized representation of metadata templates together with Web-based services to store, search, and share these templates. Templates created using CEDAR technology are stored in our openly accessible community repository. Researchers access the repository to search for appropriate templates to annotate their studies. Web-based interfaces and REST APIs enable access to all metadata templates, as well as to all the metadata collected using those templates.

**Metadata Template Creator and Template Editor:** Two highly interactive Web-based tools simplify the process of authoring metadata templates. The Template Creator allows users to create, search, and author metadata templates. Using interactive look-up services linked to the NCBO BioPortal (http://bioportal.bioontology.org), template authors can find terms in ontologies to annotate their templates and to restrict the values of template fields. The Template Creator automatically produces a user interface specification as it builds a template. The Metadata Editor uses this specification to generate a forms-based acquisition interface for acquiring individual metadata components.

**Intelligent Authoring:** To ease the burden of authoring high quality data sets, a recommender framework learns associations between data elements and suggests context-sensitive metadata values. The system can recommend possible values for metadata elements during the submission process. These value suggestions are based on an analysis of previous submissions in our metadata repository. To evaluate this approach, we analyzed all the publicly available metadata in the GEO data repository (http://www.ncbi.nlm.nih.gov/geo). We then developed an interactive Web-based value-recommender system that uses the results of this analysis to help drive GEO-oriented submissions.

Evaluation

Our team includes several community-based groups who are helping to develop and evaluate our current prototype system. These collaborators include (1) the BioSharing initiative (http://biosharing.org), which catalogs metadata standards for describing biomedical experiments, (2) ImmPort, a data warehouse of immunology-related datasets, and (3) the Human Immunology Project Consortium Standards Working Group, which designs new metadata templates and channels experimental datasets to the ImmPort repository. We successfully have represented metadata from several hundred studies provided by these groups within the CEDAR workbench. We also are working with the LINCS project (http://www.lincsproject.org) to develop a more robust metadata management pipeline that supports the authoring of metadata for a wide range of studies. Collaborations with other scientific consortia are in the planning stage, with the long-term goal of making all biomedical data easier to find, access, integrate, and reuse. ²

Acknowledgments

CEDAR is supported by NIAID grant U54 AI117925 through funds provided by the trans-NIH BD2K initiative.

References

VA’s New Electronic Health Management Platform (eHMP): Novel Features for Activity Management and Notifications

Jonathan R. Nebeker MS MD,1,2 Shane McNamee MD,1 David Douglas MD,1 James Hellewell MD,1,2 Kristian Johnson RN,1 Jessica Murphy MD,1 Ron Moody MD,3 Charlene Weir RN PhD1,2
1-Deptment of Veterans Affairs, Washington DC; 2-University of Utah, Salt Lake City UT; 3-Accenture Federal Services, Chantilly VA.

Abstract

VA has just completed the first read-write release of the Enterprise Health Management Platform (eHMP), its next generation EHR system. eHMP builds upon VA’s VistA EHR system with a data aggregation layer using JavaScript Object Notation (JSON)/Fast Healthcare Information Resources (FHIR), a service layer using modern open-source components, and a user interface layer that allows addition of compliant applications from outside VA. eHMP aggregates and standardizes data from 130 VistA instances, the Department of Defense, and community partners. The software development kit for the user-experience framework and the resource development kit for the data aggregation layer allow incorporation of applets made by others. This demonstration highlights two novel features: Activity Management and Notification services.

Description

Activity Management service allows support for clinical-process standardization. We combine business-process-management, decision-support, publish-subscribe, complex-event-processing, and team-management services to manage long running clinical processes. The initial use cases are simple tasks and consult requests. Simple tasks follow the process (but not web services) aspects of the OASIS standard for human tasks. Providing much more robust handling than consumer task-management software, this standard allows closed-loop management of task states and handling of failures through escalation or notifications back to requestor. The second use case is consult requests, which uses business process management notation and OpenCDS rules to describe and execute a process from initiation, pre-requisite validation, scheduling (including external referrals), and completion of the initial evaluation. This capability intentionally does not have backwards compatibility with CPRS/VistA and, thus, can address severe limitations of that system with communication, tracking, and escalation.

Notification service is based on the HL7 working standard for Unified Communication Service with significant extensions. The most notable extension is the handling of 10-levels of priority as generalized methods of salience varying from blocking or distraction from current workflow to subtle emphasis of information within a workflow. Levels of salience also incorporate a dimension of whether a task requires completion to resolve the notification. This notification capability is also not backwards compatible with CPRS/VistA to allow for superior alert and notification processing in eHMP. VA will gradually turn off alerts in CPRS/VistA and selectively re-implement them in eHMP to allow for appropriate presentation and automatic resolution, suppression, and escalation.

By the time of the presentation, read-write version will be deployed in a limited number of sites. At the time of submission features are in final stages of development.
iCONCUR: informed CONsent for Clinical data and biosample Use for Research

Lucila Ohno-Machado, MD, MBA, PhD, Hyoneui Kim RN, MPH, PhD, Elizabeth Bell, MPH, Xiaqian Jiang, PhD, Dexter Friedman, Claudiu Farcas, PhD
Health System Department of Biomedical Informatics, UC San Diego, La Jolla, CA

Abstract
The current practice of obtaining informed consent from patients for use of their health data for research largely depends on “blanket” consent, which involves reading a long text that is difficult to understand. As an alternative approach we developed a prototype system called iCONCUR, a tiered e-informed consent system that elicits patient preferences for clinical data sharing for research. We piloted it in two outpatient clinics of UC San Diego Medical Center. We will demonstrate our iCONCUR tool and to share our initial experience. Based on lessons learned from this pilot, we are currently developing a production version of the iCONCUR system.

System Description
iCONCUR was designed to support a granular level of control on data sharing by patients. It provides a way of granting data sharing to specific data items as well as to specific types of data users (Figure 1). iCONCUR was developed as a web-based system for easy access and use, and is hosted securely within the firewalls of our medical center. After signing into the system patients can indicate their data sharing preferences by unclicking the items they do not wish to share with particular types of data users. The data items listed in the taxonomy were determined based on the user survey and the analysis of data requests at UCSD. Patients can revise their sharing choices as frequently as they want to. The data sharing preference information is delivered to UCSD’s data concierge service, a clinical data retrieval service offered jointly through the Health System Department of Biomedical Informatics and the CTSA at UCSD. The data concierge provides clinical datasets to data requestors honoring the most up to date sharing preference at the time of data set generation. The detailed information on the data concierge transaction (e.g., data user, data use purpose, data items shared, and date/time of the data provision) is fed back to iCONCUR so that patients can check how their data are used for research by visiting the MyDataUse page in iCONCUR.

In the current phase of the project, we are focusing on augmenting iCONCUR (1) by improving the user interfaces for data sharing preference selection and data usage inspection and (2) by providing a means to incorporate proxy decisions on by authorized users.

We plan to implement the iCONCUR system for the entire UC San Diego Health System. We will conduct a larger scale survey of 1200 patients to elicit their data sharing preferences and to evaluate the effectiveness of web-based tiered informed consent system on helping patients make truly informed decisions on data sharing.
The MobiGuide Distributed & Personalized Patient Guidance System

Mor Peleg, PhD¹, Yuval Shahar, MD, PhD², Silvana Quaglini, PhD³

¹University of Haifa, Haifa, Israel; ²Ben Gurion University, Beer-Sheva, Israel; ³University of Pavia, Pavia, Italy

Abstract

MobiGuide is a ubiquitous, distributed and personalized decision-support system (DSS) for patients and their care providers. Its central DSS applies computer-interpretable guidelines (CIGs) to provide patient-specific recommendations by matching CIG knowledge with patient data that are stored in a personal health record (PHR) in interactive user-driven or proactive data driven manner. The PHR integrates data from hospital medical records, mobile biosensors, and recommendations and abstractions output by the DSS. CIGs are customized to consider the patients’ psycho-social context, quality of data, and patients’ preferences; shared decision making is supported via decision trees instantiated with patient utilities. The central DSS projects personalized CIG knowledge to a mobile DSS operating on the patients’ Smartphones that applies that knowledge locally. The system was evaluated with 10 atrial fibrillation (AF) and 19 gestational diabetes (GD) patients and their care providers in hospitals in Italy and Spain, respectively. The compliance of both patient groups to self monitoring was very high. For example, the GD patients’ self-measurement rate of Blood Glucose Levels [99%], was significantly [P<0.01] higher than that of a similar group of 247 patients managed by the hospital’s GD clinic [87%]. The system changed the clinicians’ diagnosis of 2 AF patients and resulted in starting insulin early for 2 GD patients. Patients indicated increased safety as the most important benefit. We will present a typical operation of MobiGuide (patient and doctor).

System purpose and innovation

Since the 1990’s, researchers have been developing decision-support systems (DSSs) based on computerized implementations of clinical practice guidelines. Such computer-interpretable guidelines (CIGs) can provide patient-specific recommendations during patient encounters, hence they have the potential to increase the impact on clinicians’ behavior. Traditionally, the users of CIG-based DSSs have been clinicians. The European project MobiGuide provides ubiquitous, distributed decision-support to patients in their normal environments as well as to their care providers.

The MobiGuide DSS goes beyond traditional DSSs in several ways. (1) Engaging patients: Patients are empowered by getting personalized alerts from the MobiGuide system, during their daily life, anytime, anywhere; (2) a novel mechanism of (a) considering non-clinical contexts in the customization of the CIG, including the system’s quality of data. The CIG is customized by secondary contexts, such as the patient living alone, and by technical contexts, such as a low battery or lack of an Internet link; (b) personalizing the customized CIG, by matching common events in the patient’s life with the customizable contexts. For example, a personal vacation event might be associated with the generic, though customized “irregular schedule” context that it induces. Personalization also involves specifying local preferences, (e.g., patient’s meal times on weekends), so that reminders could be delivered according to the patient’s schedule. Shared decision-making is also supported via decision-trees containing Markov models (e.g., to select an anti-coagulation drug), instantiated with patient’s utilities for various health states; (3) novel architecture for efficient distribution of decision support between a central DSS and a mobile DSS operating on the Smartphone. The central DSS has access to the full CIG and to all patient data, while the mobile DSS has access only to XML segments of the CIG knowledge that are relevant to the current patient state and are projected by the central DSS according to CIG monitored data patterns, to data measured by the sensors within the past several weeks and to partial PHR data; (5) Semantic data integration into a personal health record (PHR), which uses HL7’s virtual medical record (vMR) model to integrate data from hospital medical records, biosensors, abstractions found in patient data by the mobile or central DSS, and patient-specific recommendations delivered by the DSS. Storing vMR instances in the PHR supports data notification that is key for communication between system components; (6) The modular architecture allows plugging in different components (e.g., different decision-support engines like the decision-tree engine) and sensors, and uses generic UIs: one for patients and one for providers suitable for any chronic clinical domain.

The system was successfully evaluated with the help of patients and care providers in collaboration with hospitals in Spain and Italy in the domains of GD and AF. We will present the evaluation results. See web site for screen shots and publications: http://www.mobiguide-project.eu/
The Consent, Contact and Community Framework for Patient Reported Outcomes: Connecting ResearchKit to the Health System with FHIR

P.B. Pfiffner1,2,3, I. Pinyol1, M.D. Natter1,2 and K.D. Mandl1,3,4

1 Boston Children’s Hospital, Computational Health Informatics Program, Boston, MA; 2 University Hospital Zurich, Research Centre for Medical Informatics, Zürich, Switzerland; 3 Harvard Medical School, Department of Pediatrics, Boston, MA; 4 Harvard Medical School, Department of Biomedical Informatics, Boston, MA

Abstract

ResearchKit, Apple’s open source iOS framework, enables researchers to readily collect personal health data and patient reported outcomes (PRO) from patients “in the wild”, using their smartphone. ResearchKit is research backend agnostic, however, hence researchers must implement secure data serialization and transmission facilities themselves or choose a third party provider.

We present “The Consent, Contact and Community Framework for PRO” (C3-PRO), an open source toolchain that connects any ResearchKit app to the widely-used clinical research infrastructure Informatics for Integrating Biology and the Bedside (i2b2). This toolchain leverages the emerging health data standard Fast Healthcare Interoperability Resources (FHIR) for on-device concept representation, data serialization, and transmission. FHIR resources move creation of on-device surveys from code to file format and its use throughout the toolchain means individual components are readily substitutable.

Purpose

The development of mobile research apps – apps that collect PRO data from participants and from device sensors – has been greatly simplified by Apple’s ResearchKit, released in spring 2015. As demonstrated by the advent of ResearchStack, the Android counterpart to ResearchKit, tools for mobile development are constantly improving. Server side, many ResearchKit apps make use of SAGE services to manage participant accounts and to deliver data back to the researchers.

While this setup has already commoditized several parts of end-to-end research app development, there are still areas that researchers need to tackle. For researchers wanting to use a backend under their control, this involves:

- creating surveys and collecting responses, tasks performed programmatically
- encrypting PRO data before sending it over the wire
- providing robust, scalable server-side endpoints for the apps to connect to
- verifying the source of incoming PRO data server-side
- decrypting incoming PRO data
- storing incoming PRO data securely and in queryable form into a research database or health record system

With C3-PRO we have built an end-to-end, open source toolchain addressing these issues, consisting of four pieces. The toolchain leverages the FHIR standard for data representation and transmission and i2b2 for data storage and analytics.

The iOS framework translates between FHIR “Questionnaire” resources and ResearchKit’s in-memory representation of a survey. This means developers can use validated surveys in form of a FHIR resource instead of having to hand-code a survey task; a task that one has to repeat on Android with ResearchStack. The iOS framework also includes facilities for symmetric and asymmetric (public-key) encryption, reverse geocoding, HIPAA-compliant data de-identification and more.

Our server components use FHIR interfaces (or encrypted versions thereof) making any component substitutable. The front line component, our “C3-PRO Receiver”, can be run on Amazon for easy scalability. It performs OAuth2 dynamic client registration with each app instance and subsequently only accepts authenticated data. Since incoming data is encrypted, the receiver cannot inspect research data. It will then enqueue all authenticated resources for downstream consumption.

Our second server-side component, the “C3-PRO Consumer”, lives behind the institution firewall and consumes resources that have been enqueued by the receiver. This means it’s protected from the internet and only sees authenticated data. It also has access to the private key and is therefore able to decrypt incoming research data, ending up with a FHIR resource.

The third server-side component, the “i2b2 FHIR cell”, is an i2b2 plugin (cell) that accepts FHIR resources for mapping and storage into i2b2. This cell is being developed into a generic FHIR i2b2 cell, demonstrating the substitutable approach we are taking. It also performs the necessary data mapping to store PRO data meaningfully into i2b2.

Deployment

The C3-PRO toolchain is available open source, fully documented and with a permissive Apache 2 license, on GitHub (http://c3-pro.org/architecture.html). At the time of writing it was in production use by our own C Tracker hepatitis C ResearchKit study and partly by the Feverprints study, also from Boston Children’s Hospital. Here we will demonstrate a) how C3-PRO uses ResearchKit and ResearchStack to display FHIR Questionnaires on the smartphone, and b) how completion of a survey on the smartphone results in new data appearing in i2b2.
Bridging the Gap Between Public Health and Clinical Provider: The PHRASE Interoperable Platform for Improving Decision Support

Marc Tobias, MD¹, Naveen Muthu, MD¹, Robert Grundmeier, MD¹²

¹The Children’s Hospital of Philadelphia, Philadelphia, PA, USA; ²Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Problem: Currently, the public health community has a wealth of information regarding infection surveillance, chronic disease best practices, and other insight that is frequently difficult for providers to access. A common method for dissemination is via e-mail, which is not patient specific and may result in delayed diagnosis and intervention. Systematic delivery of decision support at the point of care can potentially improve clinician knowledge and reduce care variation, leading to better outcomes for patients while reducing cost.

Solution: The Public Health Risk Assessment Support Engine (PHRASE) was created to address this need for further integration between the public health community and frontline healthcare providers. It allows for the identification of a target population using patient characteristics like demographics, location, symptoms, vital signs, or any other data stored in the electronic health record (EHR) system. In addition, recommendations can be entered for each alert by public health officials to help drive decision support. To deliver these alerts to healthcare providers, we utilize software that accesses the data in the EHR. This data is cross-referenced with the public health rules in PHRASE. Patients who match the criteria have an alert displayed in real-time within the EHR with recommendations from the public health official (Figure 1). Similar solutions have been attempted in the past, but questions remain concerning architecture, interoperability, governance and usability. PHRASE differs from prior solutions in being compatible with current and future standards for interoperability. The interaction with the EHR was accomplished via two different platforms: 1) SMART on FHIR and 2) Care Assistant, which is The Children’s Hospital of Philadelphia’s homegrown integration software. The use of these platforms allows for scalability to institutions across a broad range of EHR vendors (Figure 2).

Status: PHRASE has established feasibility at one institution and is now expanding to a multi-institution pilot.

Reference:
LHC-Forms and Related Widgets for Capturing and Tuning Health Data

Ye Wang, MS¹, Paul Lynch, MS,¹ Ajay Kanduru, MS,¹ John Hook,² Lee Mericle, MS,¹ Christophe Ludet, MS,¹ Daniel J. Vreeman, PT, DPT, MSc², Clement J. McDonald, MD¹

¹National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
²Regenstrief Institute, Indianapolis, Indiana, USA

Abstract: NLM’s Lister Hill National Center for Biomedical Communications (LHC) has developed four inter-related open source, web-based, JavaScript tools for gathering and processing clinical data: 1) **LHC-Forms** is a data capture widget designed in partnership with Regenstrief Institute to produce a browser executable form from a stored form description. LHC-Forms supports HL7 data types, repeating groups of questions, survey scoring, validation checks, answer lists, and default values that can be derived from answers given to preceding questions. LHC-Forms can execute any LOINC panel as a form, and generate an HL7 v2 message of the entered data. Try it: [https://lhc-forms.lhc.nlm.nih.gov](https://lhc-forms.lhc.nlm.nih.gov). 2) **Clinical Table Search Service** is an auto-completion tool that provides an autocomplete menu to fields that take answers from large tables. It is accessed via URLs whose parameters control what table to search, which fields to return to the choice menu grid, and which fields of the selected item to store as hidden content in the input fields. Our implementation provides preconfigured access to many clinical tables: LOINC, RxTerms, ICD-10-CM, many NCBI genomics tables, COSMIC and others. Try it: [https://clin-table-search.lhc.nlm.nih.gov](https://clin-table-search.lhc.nlm.nih.gov). 3) An **LHC-Form building tool**. Try it: [https://lhc-formbuilder.lhc.nlm.nih.gov](https://lhc-formbuilder.lhc.nlm.nih.gov). 4) A JavaScript **validator and converter for UCUM units of measure**. Try it: [https://ucum-validator.lhc.nlm.nih.gov](https://ucum-validator.lhc.nlm.nih.gov).

The four modules can be used together or separately in web applications.

System Description: Purpose: To provide a set of sophisticated but easy to use open source web modules that can be used within existing web applications and provide pre-configured support for standard clinical vocabularies (LOINC, RxNorm, UCUM, SNOMED CT) and genomic identifiers. Developers can also configure arbitrary external tables, such as local patient and doctor registries, to work with these widgets. LHC-Forms and its sister modules are designed to work together in LHC-Forms--or separately with other applications that follow its interface specifications. LHC-Forms can produce HL7 v2 messages, and will produce FHIR clinical reports in the future. The Clinical Table Search Service supports the HL7 coded-value types: coded with exceptions (CWE) and coded with no exceptions (CNE). Both the Clinical Table Search Service’s auto-completer and LHC-Forms can be downloaded as pre-built packages for integration in a web-application, so that forms and auto-completers can be rendered with just a few lines of code. Current usage: 1) We are using LHC-Forms and the Clinical Table Search Service to model a v2 HL7 clinical genomics implementation guide which is up for HL7 ballot Sept 2016 (view full version at [https://lhc-forms.lhc.nlm.nih.gov](https://lhc-forms.lhc.nlm.nih.gov) by selecting the second example in the left column). 2128

![Image](https://lhc-forms.lhc.nlm.nih.gov)

**Figure 1.** Part of a structured genetics report form rendered in LHC-Forms to model the proposed HL7 v2 clinical genomics implementation guide which is up for HL7 ballot Sept 2016 (view full version at [https://lhc-forms.lhc.nlm.nih.gov](https://lhc-forms.lhc.nlm.nih.gov) by selecting the second example in the left column).

This research and development effort was supported in part by the Intramural Research Program of the National Institutes of Health.
Development of DataMed, a Data Discovery Index Prototype by bioCADDIE: Laying the Groundwork for Biomedical Data Discovery

Hua Xu, PhD\(^1\), Jeffrey S. Grethe, PhD\(^2\), Ian Fore, DPhil\(^4\), Ronald Margolis, PhD\(^4\), George Alter, PhD\(^5\), Susanna-Assunta Sansone, PhD\(^3\), Lucila Ohno-Machado, MD, PhD\(^2\) and the bioCADDIE team

\(^1\)University of Texas Health Science Center at Houston, TX; \(^2\)University of California, San Diego, CA; \(^3\)University of Oxford, UK; \(^4\)National Institutes of Health, Bethesda, MD; \(^5\)University of Michigan, Ann Arbor, MI

Abstract

bioCADDIE (biomedical and healthCAre Data Discovery Index Ecosystem) is a joint effort of biomedical researchers, informaticians, administrators, publishers and a funding agency, to facilitate the discovery of available data through the development of a Data Discovery Index (DDI). The initiative is funded as part of the Big Data to Knowledge (BD2K) program by the National Institute of Health. This demonstration session introduces the functional prototype of the DDI, DataMed, developed by the Core Technology Development Team (CDT), in conjunction with members of the community (Pilot Projects, Working Groups, and supplements to existing NIH-funded projects). DataMed currently covers an indexing pipeline and metadata unification for 23 data repositories selected from various biomedical subdomains. An advanced search engine presents these resources to end users. The web interface integrates several tools for specific research purposes that were developed by bioCADDIE-funded pilot projects. The demonstration aims at introducing DataMed to the informatics community and obtaining feedback from the audience.

Overview of the System

The increasing number of scientific data sets, many of which are often difficult to find online but if utilized properly could accelerate research creates the need for a data discovery index (DDI). The primary objective of the bioCADDIE CDT is to design and implement a DDI web application with a user-friendly interface that enables users to browse, search and get recommendations of datasets that best satisfy their specific interests, preferences, and needs. Our goal is to do for data what PubMed did for the literature. DataMed consists of 3 main components (Figure 1):

- A repository ingestion and indexing pipeline that maps disparate metadata from diverse data sources, into a unified specification provided by various working groups organized by bioCADDIE and related communities. This pipeline has an automated component that provides controlled translation and curation of metadata using special tools such as a transformation language and JSON-Path.
- A terminology server for terminological consistency that includes several specific ontologies such as MeSH, the Gene Ontology and the NCBI Taxonomy, which are mapped via UMLS integration.
- A web application based on the search engine that uses the discovery index for locating appropriate datasets from the processed set of repositories. This engine also uses the terminology server to expand queries using synonymous terms. In order to provide a better user experience, two modes have been designed offering basic and detailed (advanced search) options.

The DataMed demonstration will cover an overview of the web interface to the search engine and will be based on a few common search scenarios. The aim is to introduce the system to the audience and to interact with and directly collect feedback from end users of the system.

Current Status: DataMed is currently available for exploration at [http://datamed.org](http://datamed.org), encompassing 649,055 datasets from 23 different repositories ([https://datamed.org/repository_list.php](https://datamed.org/repository_list.php)) that include 10 different data types ([https://datamed.org/datatypes.php](https://datamed.org/datatypes.php)). The CDT is working to incorporate new repositories to expand the content of the DDI as well as on validating, testing and improving ranking algorithms to retrieve relevant personalized search results. We are also conducting user and usability testing.

**Acknowledgements.** bioCADDIE is funded by NIH U24AI117966 and has many more team members and collaborators than would fit the author list for this abstract. They are listed in our web site at [http://biocaddie.org](http://biocaddie.org).